The Dissertation Committee for Qinghua He
certifies that this is the approved version of the following dissertation:

Innovative Techniques for Industrial Process Modeling and Monitoring

Committee:

S. Joe Qin, Supervisor

Thomas F. Edgar

Gyeong S. Hwang

Glenn Y. Masada

Anthony J. Toprac

Innovative Techniques for Industrial Process Modeling and Monitoring

by

Qinghua He, B.E., M.S.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2005

Dedicated to Jin

Acknowledgments

I would like to express my deepest appreciation to my supervisor Dr. S. Joe Qin for his guidance and support in my research and preparation of this dissertation. I would also like to thank other members of my committee, Dr. Thomas Edgar, Dr. Gyeong Hwang, Dr. Glenn Masada and Dr. Anthony Toprac, for their time and assistance in helping me complete this dissertation.

I am grateful to Dr. Anthony Toprac, who gave me my first opportunity in graduate school to work on real industrial projects. I am fortunate to have worked with him and learned how to narrow the gap between academia and industry.

I would like to thank Dr. Martin Pottmann, Dr. Yiannis Dimitratos and other members at Process Dynamics and Control of DuPont Engineering for the opportunity of a summer internship and assistance in accomplishing part of this dissertation. It has been an enjoyable and valuable addition to my graduate education.

I also want to acknowledge the support from Advanced Process Control group at Spansion, LLC and Advanced Micro Devices, Inc.

The members of Process Modeling and Control group at UT have made my time spent at graduate school meaningful and enjoyable. I want to recognize their help and wish them all the best of luck in completing their graduate

work and excelling in their professional careers.

Most importantly I would like to thank my wife Jin for everything. I would not have been able to finish this dissertation without the love, encouragement, and support (both emotionally and technically) from her. She has been sharing the pressures and joys with me for the last five years and I cannot possibly thank her enough.

Qinghua He

Austin, Texas

March 23, 2005

Innovative Techniques for Industrial Process Modeling and Monitoring

Publication No. _____

Qinghua He, Ph.D.
The University of Texas at Austin, 2005

Supervisor: S. Joe Qin

This research presents several innovations in industrial process modeling and monitoring with the purpose of better controlling the process.

In semiconductor manufacturing industry, people show increased interested in thermal modeling of Low-Pressure Chemical Vapor Deposition (LPCVD) processes in order to understand the process better and get tighter control of film uniformity. Research in this area has resulted in several first-principles models. However, the common drawback of these models is that they are more academically-oriented than industrial-oriented, i.e., the intensive computation makes them very difficult to be applied online in order to meet the high-volume manufacturing needs. In this dissertation, a first principles transformed linear model is developed for the LPCVD process to address drawbacks of existing thermal models and facilitate its industrial implementation. The proposed model accurately predicts wafer temperatures using the

furnace wall temperatures, and it can be solved using a direct algorithm in only a few seconds. The simplicity of the model form and the fast algorithm make the model desirable for real-time updating and control of industrial scale furnaces.

In process industry, many control loops perform poorly due to reasons such as bad tuning or equipment problems. Among them, valve stiction is one of the most common equipment problems. Although there has been many attempts to understand and model valve stiction, those models are either physical models which are not practical to use, or empirical models but with rather complicated logic which make them difficult to understand and implement. In this work, a new valve stiction model is proposed with simple structure and straightforward logic which make it easy to implement. Furthermore, several published valve stiction detection techniques are reviewed. The inconsistency of Horch's first method is theoretically analyzed and illustrated by a simulated example. A new valve stiction detection method is proposed based on curve-fitting for both self-regulating and integrating processes. The new method shows superior performance to other existing methods.

Fault diagnosis plays an important role in supervision and maintenance of chemical processes in an effort to isolate the root cause once a fault is detected. The well known fault diagnosis approaches, *i.e.,* contribution plots based on Principal Component Analysis (PCA) and Partial Least Squares (PLS) models, may not explicitly identify the cause of an abnormal condition, and sometimes may lead to incorrect conclusions. In this work, a new

fault diagnosis method using fault directions in Fisher Discriminant Analysis (FDA) is developed in attempt to provide a better solution than the traditional contribution plot based on PCA. Besides, a new process monitoring method is proposed which consists of data pre-analysis, fault visualization and fault diagnosis. In both simulation example and a film industrial example, the contribution plots proposed based on fault directions in pair-wise FDA shows superior capability for fault diagnosis to the contribution plots method based on PCA.

In today's chemical industry, massive amount of data are easily available in computer controlled processes. But at the same time, the visualization of high dimensional data has been difficult. The dramatically increased computing power has not been utilized to improve the situation in industry. In this work, the commonly used visualization techniques, usually applied to relatively small static systems, are evaluated in the context of large dynamic systems. A general framework of hierarchical visualization is proposed and several multivariate visualization methods are developed in this work. The performance of PCA, PLS, Class Preserving Projection (CPP), FDA and two proposed approaches based on Support Vector Machines (SVM) are compared using an industrial data set.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction and Dissertation Outline

Process control and process monitoring as the two complementary parts of modern control systems have gone through tremendous development in the past a few decades and become more and more popular in industrial applications. Arguably, mathematical system theory is one of the most significant achievements of twentieth-century science, but its practical impact is only as important as the benefits it can bring [33] and there are many gaps between theories and practical applications. This dissertation has been trying to fill some of these gaps.

Since Kalman introduced the state-space representation and laid the foundation for state-space based optimal filtering and optimal control theory around 1960, different types of model-based control design techniques have been developed, with linear quadratic (LQ) optimal control as the cornerstone [32]. As revealed by its name, process model is the foundation of the model-based control methods. Although adaptive control, robust control techniques are developed to address the issue of process-model mismatch, a model structure or an initial model which is within certain uncertainty range is still required. With the application of the advanced control techniques to large

and complex systems, such as microelectronics manufacturing processes and pharmaceutical processes, people often find that lack of a feasible model is usually the "bottleneck" for control performance of industrial processes, and believe that improving understanding of the process and building fundamental process models for use in control would have a much larger impact on controlling the process than would efforts focused on improving control parameter optimization techniques or incorporating more complex algorithms to control [9]. However, even for the cases where accurate models are available, there are times that the required computation is too intensive for online application, and other times that algorithms are so complicated that implementation becomes a hinderance. This is the case for the thermal modeling of low pressure chemical vapor deposition (LPCVD) furnace where most existing models take hours to get a converged solution and the models consist of hundreds of partial and ordinary differential equations. My work in this subject resulted in a first principles transformed linear model which takes seconds to solve with a direct algorithm [40–43]. Another example is the modeling of valve stiction where physical models are not practical because of the difficulty of obtaining some of the model parameters. Although there are a couple of empirical data-driven models developed recently, their model structure and logic are rather complicated, making them difficult to implement. This issue is addressed by the proposed new valve stiction model in this work [44, 45]. Furthermore, based on the understanding of the characteristics of valve stiction, a curve fitting method is developed to detect valve stiction for both

self-regulating and integrating processes.

For process monitoring, up until the late 1980′s, traditional fault detection and identification methods were based on a mathematical model of the system. Since then, the process control community began to investigate the use of multivariate statistics for process monitoring, and statistical process monitoring (SPM) has become one of the most active research areas in last decade [80]. To overcome the shortcomings of univariate Statistical Process Control (SPC) charts, multivariate statistical methods, such as principal component analysis (PCA) and partial least squares (PLS), have been applied to generate SPC charts for fault detection. Because of the data-based nature of the SPM methods, it is relatively easy to apply to processes of rather large scale comparing to other methods based on systems theory or rigorous process models. Today, SPM has found wide applications in different industrial processes, especially for fault detection. After a fault has been detected, it is highly desirable to isolate the root cause of the fault, especially for complex process which has very large number of variables. However, the well known fault diagnosis approaches, *i.e.,* contribution plots based on PCA and PLS models, may not explicitly identify the cause of an abnormal condition, and sometimes may lead to incorrect conclusions. In this work, a new fault diagnosis method using fault directions in Fisher discriminant analysis is developed in order to provide a better solution than the traditional contribution plot based on PCA [46–48]. On the other hand, the visualization of the dynamic behavior of complex processes using high dimensional data has been quite difficult and

largely unsolved issue, and not much effort has appeared in this area. In this dissertation, a general framework of hierarchical visualization is proposed and several multivariate visualization methods are developed for SPM [38, 39].

To summarize, this dissertation focuses on developing industrial process modeling and monitoring techniques with four major parts:

In Chapter 2, a new thermal model is developed to predict wafer temperatures within a hot-wall Low Pressure Chemical Vapor Deposition (LPCVD) furnace using the furnace wall temperatures as measured by thermocouples. Model predictions show excellent agreement with experimental data. Based on an energy balance of the furnace system, this model is a transformed linear model which captures the nonlinear relationship between the furnace wall temperature distribution and the wafer temperature distribution. The model can be solved with a direct algorithm instead of iterative algorithms which are used in all other existing thermal models. Since the direct algorithm is non-iterative, there is no convergence problem, nor local minima problem related to nonlinear optimization. In addition, the direct algorithm greatly reduces the computation effort. Configuration factors are calculated by a finite area to finite area method, which avoids numerical integration methods that are much more difficult to implement and require more computation. The simplicity of the model form and the fast algorithm make the model desirable for real-time updating and control.

Chapter 3 reviews several published valve stiction models and presents a new valve stiction model which has a simple structure and straightforward

logic. Furthermore, several published valve stiction detection techniques are reviewed. The inconsistency of Horch's first method is theoretically analyzed and illustrated by a simulated example. A new valve stiction detection method is proposed and its theoretical analysis is presented. Stiction index (SI) is introduced based on the proposed method to facilitate the automation of the method. Superior performance of the proposed method is demonstrated using both simulated and industrial examples.

Chapter 4 presents a new process monitoring method which is composed of three parts: (i) a pre-analysis step that first roughly identifies various clusters in a historical data set and then precisely isolates normal and abnormal data clusters by the k-means clustering method; (ii) a fault visualization step that visualizes high-dimensional data in 2-D space by performing global Fisher discriminant analysis (FDA), and (iii) a new fault diagnosis method based on fault directions in pair-wise FDA. A simulation example is used to demonstrate the performance of the proposed fault diagnosis method. An industrial film process is used to illustrate a realistic scenario for data pre-analysis, fault visualization and fault diagnosis. In both examples, the contribution plots method based on fault directions in pair-wise FDA shows superior capability for fault diagnosis to the contribution plots method based on PCA.

Multivariate visualization techniques have been developed in the fields of statistics, artificial intelligence and computer graphics and have been widely used in these fields as fundamental tools to allow human eyes to detect special structures in data. In Chapter 5, the commonly used visualization techniques,

usually applied to relatively small static systems, are evaluated in the context of large dynamic systems. In general, people are interested in the visualization of the static properties (such as outliers, clusters and variable correlations) and dynamic properties (such as process drifts, shifts and oscillations) which can be visualized either in the original variable space or in the transformed space. For the visualization of dynamic properties in the high dimensional original space, Dynamic Parallel Coordinates (DPC) is proposed; variable grouping is introduced to reduce clutter in handling large data sets and hierarchical visualization scheme is proposed to provide a general framework for visualization and exploration of large multivariate data sets. For class visualization in the transformed space, principal component analysis (PCA), partial least squares (PLS) and class-preserving projection (CPP) are evaluated. It is demonstrated that some commonly used classification methods such as Fisher discriminant analysis (FDA) and support vector machines (SVM) can be tailored for high dimension class visualization. A binary-tree approach and a cross-selection approach are proposed based on SVM. The performance of PCA, PLS, CPP, FDA and two approaches based on SVM are compared using an industrial data set.

The last part of this dissertation summarizes the major contributions of this work and gives suggestions on future directions.

Chapter 2

Computationally Efficient Modeling of Wafer Temperatures in an LPCVD Furnace

2.1 Introduction

Chemical Vapor Deposition (CVD) is one of several film deposition techniques which are used extensively in the fabrication of microelectronics devices. CVD has become extremely popular and is the preferred deposition method for a wide range of materials [10], especially for the deposition of insulating and semiconducting films. Compared to other film formation methods, CVD offers excellent control of film structure and composition, reasonable deposition rates, and good step coverage [49]. Step coverage is a particular concern with submicron technologies where very small contacts are needed for the coverage of high aspect ratio features. Low Pressure CVD (LPCVD) reactors deposit polycrystalline and amorphous films at moderate temperature (400 to 650 ^oC) and low pressure (0.2 to 2 Torr). These films are deposited on the epitaxial substrate and patterned to form various circuit structures. LPCVD reactors can be divided into hot and cold wall systems. Hot-wall systems have the advantages of uniform temperature distribution and high throughput—more than one hundred wafers can be deposited simultaneously. Virtually all polycrystalline silicon and a considerable amount of dielectric

deposition are done in hot-wall systems.

The uniformity and deposition rate are key factors for the successful operation of LPCVD and can be controlled by manipulating the heat and mass transfer process occurring in the reactor. To achieve reasonable deposition uniformity, hot-wall multiwafer LPCVD reactors are usually operated in the reaction rate limited region, requiring excellent temperature control and temperature uniformity. Gas flow dynamics are negligible in this situation. Badgwell et al. [4] show that wafer-to-wafer deviations in average growth rate correlate very well with wafer-to-wafer deviations in average wafer temperature. It appears that the key to improving film uniformity is to provide the wafers with a uniform thermal environment. In this sense, thermal modeling plays a key role in the modeling and control of the hot-wall LPCVD process. The focus of this work is on the thermal modeling of hot-wall LPCVD.

A number of hot-wall multiwafer LPCVD thermal models have been developed, some of which are shown in Table 2.1.

Table 2.1: Published thermal models in hot-wall LPCVD

First Author	Year	Type [1]	Solution [2]	Dimension [3]	Data [4]
Matsuba [73]	1985	N	N	2	Yes
Van Schravendijk [87]	1987	N	N	1	No
Tavel [91]	1988	N	N	1	Yes
Hirasawa [51]	1989	N	N	1	Yes
De Waard [96]	1992	N	N	1	Yes
Houf [55]	1993	N	N	2	Yes
Hirasawa [50]	1993	N	N	2	Yes
Badgwell [5]	1994	N	N	2	Yes
Coronell [17]	1994	M	N	2	Yes
Azzaro [3]	1995	N	N	1	Yes
Kim [65]	1999	N	N	2	Yes
Park [77]	2000	N	N	2	Yes
This work	2002	N	A	1	Yes

[1] Model types are Nonlinear (N), Monte Carlo simulation (M).
[2] Solutions are Analytical (A), Numerical (N).
[3] Solution dimensions are: 1, 2 or 3.
[4] Model predictions were compared to experimental data (Yes) or (No).

Sato [86] measured the emissivity of silicon in the spectral region from 0.4 to 15μm at various temperatures from 70 to 800^oC. Those data were used as references by other authors in their thermal models [3, 5]. Hu [57] proposed a model of radiative transfer in which transient temperature profiles were analyzed in a row of wafers during cooling. Convective heat transfer was shown to be entirely negligible in comparison with radiative heat transfer. Although Hu's model is not a complete thermal model for CVD reactors, his radiative

modeling approach has been widely adopted by others. Matsuba et al. [73] modified Hu's model to include energy balances for the outer tube and boat. Radiative heat transfer was allowed between the wafers, tube wall, and the boat. Van Schravendijk and De Koning [87] developed a radiative heat transfer model within the furnace which included energy balances for the insulation, reactor doors, heating coils, process tube and wafers. Radial wafer temperature gradients were ignored, diffusive emission and reflection were assumed and the wafer load was approximated as a cylinder with an "effective" heat conductivity. Tavel and Hearn [91] developed a computer simulation program describing the heating and cooling of a row of silicon wafers undergoing a prescribed thermal cycle. Unlike Hu's work where only radiation was considered, their simulation also took the effects of conduction and convection into account. Hirasawa and Takagaki [51] developed a thermal model based on an energy balance for the wafers, combined with equations describing radiation to the tube walls. De Waard and De Koning [96] revised Van Schravendijk and De Koning's model by admitting direct radiation from the heating element to the wafers and including heat transfer by thermocouple sheaths. Houf et al. [55] developed a transient two-dimensional model with the simplification that interior wafers in the load were allowed to exchange radiation with only the portion of the tube wall located directly above them. Badgwell et al. [5] developed a new energy balance model to predict wafer temperatures in a hot-wall multiwafer LPCVD reactor. The model is most similar to the radiation models of Hirasawa and Takagaki [51] and De Waard and De Koning [96],

with extensions to include more realistic geometry and radial heat transfer within the wafers. The model predictions compared favorably with *in situ* wafer temperature measurements described in a related paper [4]. Coronell and Jensen [17] provided an alternative approach based on a direct simulation Monte Carlo technique to simulate the radiation heat transfer in a multiwafer LPCVD reactor. Azzaro and Couderc [3] presented a thermal model involving computations of the wafer temperature and taking into account radiative exchanges occurring inside an LPCVD reactor with non-isothermal tube walls. Recently, Kim and Kim [65] and Park et al. [77] presented analyses of heat transfer in a LPCVD reactor with thermal models very similar to Badgwell's except that specular reflection was considered in [77].

All these models mentioned above involve nonlinear optimization and other advanced numerical techniques such as hybrid Newton-time integration [55], trapezoidal numerical integration method [50], and sparse-matrix techniques [5], and thus are computationally demanding. For instance, Coronell and Jensen's [17] Monte Carlo simulation required 6 hours of computation on an IBM RS-6000/350 workstation for a commercial scale BTU/Bruce LPCVD reactor. In this work, we develop a new model which captures the nonlinear relationship between the furnace wall temperature distribution and the wafer temperature distribution to provide accurate prediction of wafer temperatures. The computation takes several seconds on a personal computer or laptop for the same LPCVD system mentioned above. With the simple linear structure, the model can be conveniently integrated into model-based control

algorithms, such as run-to-run control. Instead of using iterative algorithms, the new model uses a direct algorithm, greatly reducing computation effort.

The remaining part of the chapter is organized as follows. In Section 2.2, we give some background information about the LPCVD system, the geometry of the furnace system we used in this work, notations, numbering, and model assumptions. Section 2.3 presents the detailed model based on an energy balance within the LPCVD reactor. Model verification and sensitivity analysis are presented in Section 2.4. Section 2.5 gives conclusions.

2.2 Preliminaries

In this section, the geometry of the furnace system used in this work is first introduced, notation and numbering are explained, and model assumptions discussed.

2.2.1 Apparatus

The model we present here is applicable to both horizontal and vertical furnaces. In order to compare model predicted wafer temperatures with Badgwell's *in situ* measured wafer temperatures, we use the same reactor (BTU/Bruce furnace) geometry as described in [4]. Fig. 2.1(a) shows the schematic configuration of an industrial scale BTU/Bruce hot-wall multiwafer LPCVD furnace. Six-inch wafers are arranged vertically and concentrically inside the quartz tube and heated by five independently controlled heating elements wrapped around the quartz tube. An elastomeric O-ring is used for

door seals. Consequently, both ends of the reactor are kept cooler than O-ring operation temperature (about 250^oC) by water cooling. This makes the thermal environments different for different zones in the furnace. For convenience, the furnace wall is conceptually divided into three zones: inlet zone, heating zone and outlet zone as shown in Fig. 2.1. The inlet zone is from the front door to the beginning of the first heating element, the heating zone from the beginning of the first heating element to the end of the fifth heating element, and the outlet zone from the end of the fifth heating element to the back door. Dimension and other specification of the reactor used in this work are listed in Table 2.2.

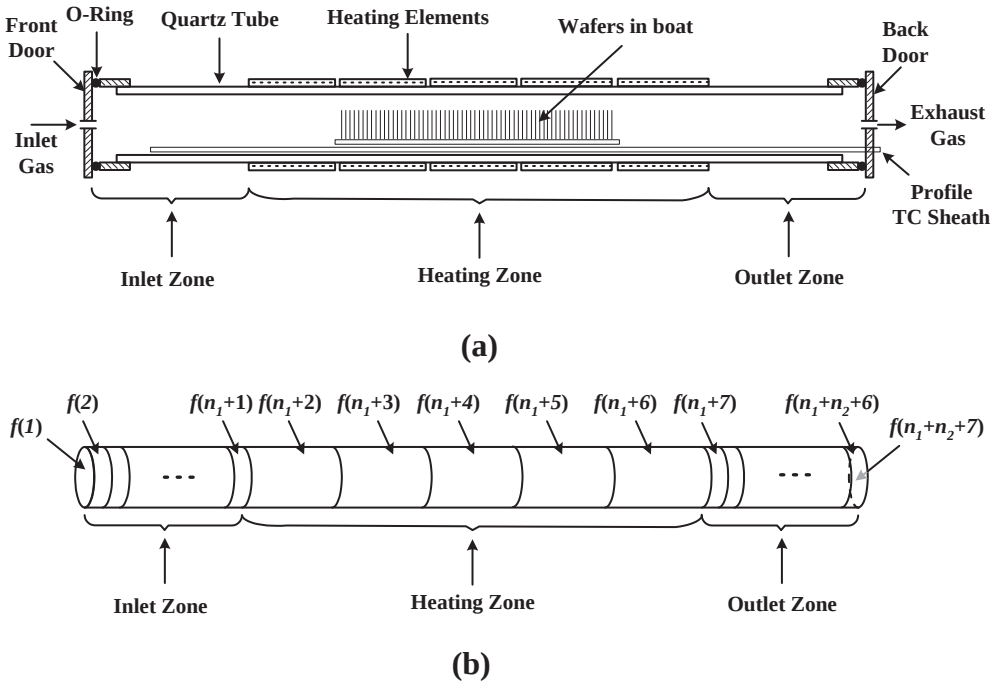


Figure 2.1: (a) Schematic drawing of BTU/Bruce LPCVD furnace, (b) Discretized furnace elements

Table 2.2: SEMATECH BTU/Bruce furnace geometry [6]

Parameter	SEMATECH Reactor
Reactor Length L (m)	2.29
First Wafer Location [5](m)	0.786
Last Wafer Location (m)	1.50
Interwafer Spacing t (m)	0.00479
Furnace Inner Radius (m)	0.145
Wafer Radius r_w (m)	0.075
Wafer Thickness t_w (mm)	0.7
Number of wafers m	150
Front and back door temperatures T_d (^oC)	250

[5] Locations are given relative to the front door of the reactor.

In this manufacturing system, the wafer temperatures can not be measured directly; only the tube wall temperatures are directly measured by thermocouples. Therefore, our model aims to relate tube wall temperatures to wafer surface temperatures based on an energy balance among wafers, furnace tube walls and furnace doors. For computational convenience, the furnace is discretized into (n_1+n_2+7) furnace elements, $f(1), f(2), \cdots, f(n_1+n_2+7)$, as shown in Fig. 2.1(b). The front door is represented by $f(1)$. The inlet zone is discretized into n_1 equal or non-equal length elements: $f(2), f(3), \cdots, f(n_1+1)$. Five furnace elements $f(n_1+2), f(n_1+3), \cdots, f(n_1+6)$ corresponds to the five heating elements. The outlet zone is discretized into n_2 equal or non-equal length elements: $f(n_1+7), f(n_1+8), \cdots, f(n_1+n_2+6)$. The last element $f(n_1+n_2+7)$ is the back door. Each furnace element $f(i)$ is assumed to be

an isothermal disk or cylinder. There are m wafers inside the furnace. For reference, wafers are numbered from left to right in similar manner to the numbering of furnace elements. Each wafer is assumed to be an isothermal disk. The modeled system therefore consists of quartz tube, wafers, front door and back door, with a total number of $(m + n_1 + n_2 + 7)$ discretized elements for this closed system.

2.2.2 Model assumptions

The basic assumptions needed in this work are listed below. They are standard in thermal modeling, and have been widely used by other researchers.

1. The reactor is axially symmetric.

   The wafer boat, profile thermocouple sheaths and gas injectors, due to their small geometry, are assumed to not interfere with heat transfer.

2. Heat conduction and convection of the gas phase are negligible.

   This hypothesis has been theoretically analyzed [3, 57] and confirmed by experiments [4]. Therefore, the only significant mode of heat transfer, which is also the only mode we considered in our model, is direct radiation. This hypothesis also indicates that heat transfer does not depend on mass transfer and thus the energy balance can be solved separately from the mass balance. This underlines the common approach of modeling thermal effects independently from chemical reaction mechanisms and kinetics in LPCVD processes.

3. The radial temperature gradient within the wafer is negligible.

   Although many researchers presented theoretical and experimental evidence for wafer radial temperature variations [4, 50, 100], the variations are expected to be small for all the wafers except the end wafers, and thus have been neglected by many researchers [3, 17, 83, 87, 96]. In our case, since the wafer-to-wafer temperature uniformity is the only interesting aspect, this assumption will reduce the degree of complexity significantly without adding severe restriction to the model.

4. All the surfaces are diffuse-gray surfaces which emit and reflect radiation diffusely with constant emissivity and reflectivity.

   Interior surfaces of the quartz tube are assumed to be coated with the same film as wafer surfaces and thus have the same value of emissivity. Two different type of reflections have been used in thermal modeling: 1) specular reflection, in which the angle of reflection is equal to the angle of incidence, and 2) diffuse reflection, in which incident radiation is reflected equally in all directions. Whether reflection is specular or diffuse depends on the optical roughness, defined as the ratio of the root-mean-square roughness height of the surface to the wavelength of the radiation. Specular reflection occurs when the optical roughness is much less than unity [88]. At shorter wavelengths, or for LPCVD films possessing a certain degree of roughness (i.e., polysilicon films), the photon is reflected in a diffuse manner [17]. De Waard and De Koning [96] state that "bare silicon wafers tend to exhibit mirror-like reflections, but oxidated wafers do

not". Badgwell et al. [5] argued that the assumption of diffuse reflection would not be expected to cause serious errors in the model even if the polysilicon surfaces do in fact reflect specularly within the reactor. This is verified by Coronell and Jensen [17] in their direct Monte Carlo simulation even for the highly polished Si wafer. The comparison of wafer temperature profiles for specular and diffuse reflections shows that there is no significant difference between the two cases. There is no agreement among researchers on which one is better, nor are there clear criteria for judgment. Researchers tend to use the one which can either facilitate the analysis or simplify the computation [5, 57, 96]. In our case, because of the roughness of the polysilicon film, diffuse reflection is a reasonable assumption.

5. The gases in the reactor are nonparticipating medium.

   Such medium neither emits, absorbs, nor scatters. It has no effect on the transfer of radiation between surfaces. A vacuum meets these requirements exactly, along with most gases to excellent approximation [88].

6. The reactor doors are isothermal disks of stainless steel.

   Their emissivity is 0.37 according to Siegel and Howell [88].

Steady-state modeling is considered here since we are interested in the deposition rate and uniformity which are affected mainly by steady-state temperatures in the furnace. Transient models would be useful to describe the

thermal stresses, which are beyond our interest here. The main purpose of the steady-state modeling is to perform run-to-run control of wafer temperature uniformity and thus film thickness uniformity.

2.3 Modeling of the Hot-Wall Multiwafer LPCVD Reactor

As illustrated in Fig. 2.1 (b), the wafer surfaces, the interior surface of the front door, the interior surface of the quartz tube, and the interior surface of the back door consist of an enclosure of $m + n_1 + n_2 + 7$ discrete surface areas. For simplicity, we define $n \equiv n_1 + n_2 + 7$ and $N \equiv m + n$. All these N surfaces are diffuse-gray surfaces according to assumption 4. Considering the energy balance of the kth surface area A_k of the enclosure at its steady-state temperature, we have

$$q_k A_k = (q_{o,k} - q_{i,k}) A_k \tag{2.1}$$

where A_k is the area of surface k, $q_{i,k}$ and $q_{o,k}$ are the rates of incoming and outgoing radiative energy per unit area of surface k. Therefore, $(q_{o,k} - q_{i,k})A_k$ represents the net radiative energy loss of surface k. q_k is the energy flux supplied to the surface k by some means other than the radiation inside the enclosure to balance the net radiative energy loss and thereby maintain the specified steady-state surface temperature.

A second equation results from the fact that the energy flux leaving the surface k is composed of the energy emitted by surface k plus the energy

reflected by surface k. This gives

$$q_{o,k} = \epsilon_k \sigma T_k^4 + \rho_k q_{i,k} = \epsilon_k \sigma T_k^4 + (1 - \epsilon_k) q_{i,k} \tag{2.2}$$

where ϵ_k is the emissivity of the surface k, σ is Stefan-Boltzmann constant $5.67051 \times 10^{-8} W/(m^2 \cdot K^4)$, T_k is the temperature of the surface k in Kelvin, and ρ_k is the reflectivity of the surface k. Here we make use of the fact that $\epsilon + \rho = 1$ for a diffuse-gray surface. We take 0.67 as the emissivity value for both furnace interior surface and wafer surfaces and treat it as constant within the temperature range from 600^oC to 630^oC. This is consistent with Sato's emissivity measurement of silicon at 600^oC which is approximately 0.65. The front and back doors are made of stainless steel with emissivity 0.37 estimated from Siegel and Howell [88].

The incident energy of surface k is the summation of the portions of the energy leaving all the surfaces in the enclosure that arrive at surface k:

$$A_k q_{i,k} = A_1 F_{1-k} q_{o,1} + A_2 F_{2-k} q_{o,2} + \cdots + A_j F_{j-k} q_{o,j} + \cdots + A_k F_{k-k} q_{o,k} + \cdots + A_N F_{N-k} q_{o,N} \tag{2.3}$$

where F_{j-k} is the configuration factor from finite area j to finite area k which defines the fraction of energy leaving surface j that arrives at surface k. If the kth surface is concave, a portion of its outgoing flux will contribute directly to the incident flux.

For configuration factors between finite areas, we have the reciprocity relation:

$$A_1 F_{1-2} = A_2 F_{2-1} \tag{2.4}$$

From this relation, we have

$$\begin{aligned} A_1 F_{1-k} &= A_k F_{k-1} \\ A_2 F_{2-k} &= A_k F_{k-2} \\ &\vdots \\ A_N F_{N-k} &= A_k F_{k-N} \end{aligned} \tag{2.5}$$

Rewriting (2.3) by replacing all areas with A_k, we get:

$$A_k q_{i,k} = A_k F_{k-1} q_{o,1} + A_k F_{k-2} q_{o,2} + \cdots + A_k F_{k-j} q_{o,j} + \cdots + A_k F_{k-k} q_{o,k} + \cdots + A_k F_{k-N} q_{o,N} \tag{2.6}$$

Eliminating A_k from both sides, we get the incident flux:

$$q_{i,k} = \sum_{j=1}^{N} F_{k-j} q_{o,j} \tag{2.7}$$

Substitute (2.2) into(2.1) to eliminate $q_{i,k}$:

$$q_k = \frac{\epsilon_k}{1-\epsilon_k} \left(\sigma T_k^4 - q_{o,k} \right) \tag{2.8}$$

This is the energy balance for surface k in terms of its temperature and outgoing energy flux.

Substitute (2.7) into (2.1) to eliminate $q_{i,k}$:

$$q_k = q_{o,k} - \sum_{j=1}^{N} F_{k-j} q_{o,j} = \sum_{j=1}^{N} F_{k-j} \left(q_{o,k} - q_{o,j} \right) \tag{2.9}$$

This is the energy balance for surface k in terms of outgoing energy flux for every surface in the enclosure.

From (2.8) we have

$$q_{o,k} = \sigma T_k^4 - \frac{1-\epsilon_k}{\epsilon_k} q_k \tag{2.10}$$

and

$$q_{o,j} = \sigma T_j^4 - \frac{1-\epsilon_j}{\epsilon_j} q_j \tag{2.11}$$

Substituting (2.10) and (2.11) into (2.9) gives:

$$q_k = \sum_{j=1}^{N} F_{k-j} \left(\sigma T_k^4 - \frac{1-\epsilon_k}{\epsilon_k} q_k - \sigma T_j^4 + \frac{1-\epsilon_j}{\epsilon_j} q_j \right) \tag{2.12}$$

or

$$q_k = \sigma T_k^4 \sum_{j=1}^{N} F_{k-j} - \frac{1-\epsilon_k}{\epsilon_k} q_k \sum_{j=1}^{N} F_{k-j} - \sum_{j=1}^{N} \sigma F_{k-j} T_j^4 + \sum_{j=1}^{N} \frac{1-\epsilon_j}{\epsilon_j} q_j F_{k-j} \tag{2.13}$$

Note that for an enclosure:

$$\sum_{j=1}^{N} F_{k-j} = 1 \tag{2.14}$$

We therefore have

$$q_k = \sigma T_k^4 - \frac{1-\epsilon_k}{\epsilon_k} q_k - \sum_{j=1}^{N} \sigma F_{k-j} T_j^4 + \sum_{j=1}^{N} \frac{1-\epsilon_j}{\epsilon_j} q_j F_{k-j} \tag{2.15}$$

or

$$\frac{q_k}{\epsilon_k} - \sum_{j=1}^{N} \frac{1-\epsilon_j}{\epsilon_j} q_j F_{k-j} = \sigma T_k^4 - \sum_{j=1}^{N} \sigma F_{k-j} T_j^4 \tag{2.16}$$

The next step is the analysis of the radiation exchange between the surface areas. Two types of boundary conditions are involved: (1) for the furnace elements, the required energy supplied to the surface must be determined

given a specified surface temperature, and (2) for the wafers, the temperature must be determined when a known heat input is imposed.

Since we assume that each wafer is an isothermal disk, there is no net heat transfer within the wafer by conduction, nor any heat transfer mode other than radiation between wafers and other surfaces. At steady-state, we must have $q_{o,k} = q_{i,k}$ and $q_k = 0$ for $k = 1, 2, \cdots, m$. Notice that the wafer temperature T_k $(k = 1, 2, \cdots, m)$ is unknown and must be determined. Therefore, equation (2.16) is simplified as:

$$\sum_{j=1}^{n} \frac{1 - \epsilon_{fj}}{\epsilon_{fj}} q_{fj} F_{wk-fj} = -\sigma T_{wk}^4 + \sum_{j=1}^{m} \sigma F_{wk-wj} T_{wj}^4 + \sum_{j=1}^{n} \sigma F_{wk-fj} T_{fj}^4 \quad (2.17)$$

for $k = 1, 2, \cdots, m$.

To distinguish wafers from furnace elements, we use subscript w to denote variables for wafers, and subscript f to denote variables for furnace elements. ϵ_{fj} is the emissivity of the jth furnace element, q_{fj} is the energy flux provided to the jth furnace element by surroundings outside the enclosure, F_{wk-fj} is the configuration factor from the kth wafer to the jth furnace element, T_{wk} is the temperature of the kth wafer, F_{wk-wj} is the configuration factor from the kth wafer to the jth wafer, and T_{fj} is the temperature of the jth furnace element.

For furnace element surface areas, temperatures are assumed to be known while energy supplied by means outside the enclosure is unknown. Equation (2.16) is rewritten as:

$$\frac{q_{fk}}{\epsilon_{fk}} - \sum_{j=1}^{n} \frac{1-\epsilon_{fj}}{\epsilon_{fj}} q_{fj} F_{fk-fj} = \sigma T_{fk}^4 - \sum_{j=1}^{m} \sigma F_{fk-wj} T_{wj}^4 - \sum_{j=1}^{n} \sigma F_{fk-fj} T_{fj}^4 \quad (2.18)$$

for $k = 1, 2, \cdots, n$.

Considering Equations (2.17) and (2.18) together, q_f's and T_w's are N unknowns, and others are known or can be calculated. We have a total of N equations and N unknowns. We can obtain a unique solution since these equations are linearly independent.

Defining

$$E_f \equiv \begin{bmatrix} \frac{1-\epsilon_{f1}}{\epsilon_{f1}} & 0 & \cdots & 0 \\ 0 & \frac{1-\epsilon_{f2}}{\epsilon_{f2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1-\epsilon_{fn}}{\epsilon_{fn}} \end{bmatrix}$$

$$E'_f \equiv \begin{bmatrix} \frac{1}{\epsilon_{f1}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\epsilon_{f2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\epsilon_{fn}} \end{bmatrix}$$

$$F_{w-f} \equiv \begin{bmatrix} F_{w1-f1} & F_{w1-f2} & \cdots & F_{w1-fn} \\ F_{w2-f1} & F_{w2-f2} & \cdots & F_{w2-fn} \\ \vdots & \vdots & \ddots & \vdots \\ F_{wm-f1} & F_{wm-f2} & \cdots & F_{wm-fn} \end{bmatrix}$$

$$F_{w-w} \equiv \begin{bmatrix} F_{w1-w1} & F_{w1-w2} & \cdots & F_{w1-wm} \\ F_{w2-w1} & F_{w2-w2} & \cdots & F_{w2-wm} \\ \vdots & \vdots & \ddots & \vdots \\ F_{wm-w1} & F_{wm-w2} & \cdots & F_{wm-wm} \end{bmatrix}$$

$$F_{f-f} \equiv \begin{bmatrix} F_{f1-f1} & F_{f1-f2} & \cdots & F_{f1-fn} \\ F_{f2-f1} & F_{f2-f2} & \cdots & F_{f2-fn} \\ \vdots & \vdots & \ddots & \vdots \\ F_{fn-f1} & F_{fn-f2} & \cdots & F_{fn-fn} \end{bmatrix}$$

$$F_{f-w} \equiv \begin{bmatrix} F_{f1-w1} & F_{f1-w2} & \cdots & F_{f1-wm} \\ F_{f2-w1} & F_{f2-w2} & \cdots & F_{f2-wm} \\ \vdots & \vdots & \ddots & \vdots \\ F_{fn-w1} & F_{fn-w2} & \cdots & F_{fn-wm} \end{bmatrix}$$

$$Q_f \equiv \begin{bmatrix} q_{f1} \\ q_{f2} \\ \vdots \\ q_{fn} \end{bmatrix} \qquad T_w^4 \equiv \begin{bmatrix} T_{w1}^4 \\ T_{w2}^4 \\ \vdots \\ T_{wm}^4 \end{bmatrix} \qquad T_f^4 \equiv \begin{bmatrix} T_{f1}^4 \\ T_{f2}^4 \\ \vdots \\ T_{fn}^4 \end{bmatrix}$$

Equations (2.17) and (2.18) can be written as:

$$F_{w-f}E_fQ_f = -\sigma T_w^4 + \sigma F_{w-w}T_w^4 + \sigma F_{w-f}T_f^4 \tag{2.19}$$

and

$$E'_fQ_f - F_{f-f}E_fQ_f = \sigma T_f^4 - \sigma F_{f-w}T_w^4 - \sigma F_{f-f}T_f^4 \tag{2.20}$$

Rearrange (2.20) to solve for Q_f:

$$Q_f = \sigma(E'_f - E_fF_{f-f})^{-1}(T_f^4 - F_{f-w}T_w^4 - F_{f-f}T_f^4) \tag{2.21}$$

Substituting (2.21) into (2.19) and rearranging the equation, we have

$$\left[I_m - F_{w-w} - F_{w-f}E_f\left(E'_f - E_fF_{f-f}\right)^{-1}F_{f-w}\right]T_w^4 =$$
$$F_{w-f}\left[I_n - E_f\left(E'_f - E_fF_{f-f}\right)^{-1}(I_n - F_{f-f})\right]T_f^4 \tag{2.22}$$

where I_m and I_n are m by m and n by n identity matrices.

Denoting

$$A_1 = \left[I_m - F_{w-w} - F_{w-f}E_f\left(E'_f - F_{f-f}E_f\right)^{-1}F_{f-w}\right]$$

$$A_2 = F_{w-f}\left[I_n - E_f\left(E'_f - F_{f-f}E_f\right)^{-1}\left(I_n - F_{f-f}\right)\right]$$

we solve (2.22) to obtain T_w^4:

$$T_w^4 = CT_f^4 \tag{2.23}$$

where $C = A_1^{-1}A_2$.

As discussed in Assumption 4, the emissivities are constants within the normal process temperature range. Therefore, A_1 and A_2 are constant coefficient matrices. They are independent of the process temperature. As a consequence, for a specified furnace device, we only need to calculate C once. The 4^{th} power of the wafer temperature (T_w^4) is just a linear combination of the 4^{th} power of the furnace element temperature (T_f^4). Given T_f, we only need a matrix-vector multiplication to get T_w. This greatly simplifies the calculation process if we want to optimize or control the wafer temperature profile by changing the furnace temperature profile. It is also very convenient for online updating.

A remaining problem is how to fill matrices A_1 and A_2 with configuration factors. Many researchers use differential configuration factor dF_{dk-dj} to give the fraction of radiation emitted from the differential surface element dk which is intercepted by the differential surface element dj, then integrate over whole area to get the total heat exchange between two finite areas. This

procedure requires numerical integration methods. In our case, where wafer temperatures must be found, the solution can require moderate to considerable effort if we use the above method. Instead of using a numerical integration method, we use published configuration factor formulae for finite area to finite area or combinations of these formulae to calculate configuration factors, greatly simplifying the procedure. A good collection of published configuration factors is given by Howell [56].

2.4 Results and Sensitivity Analysis

In the previous analysis, we have assumed that the temperatures of all n furnace elements are known. In practice, however, we only know the temperature where it is measured. In this work, we adopted the experimental furnace temperature profile used by Badgwell where only seven temperatures ($f(1), f(n_1+2), f(n_1+3), \cdots, f(n_1+6)$ and $f(n)$) were measured. In Badgwell's experiments, the powers for heating elements were controlled such that five elements in heating zone have the same temperature ($615\,^oC$). The front door and back door temperatures were controlled to $25\,^oC$ due to the application of O-ring seals as previously discussed. The exact temperature profile in the inlet and outlet zones close to the cold doors is unknown. In the following simulations and sensitivity analysis, we assume a simple linear temperature drop from the ends of the heating zone to the doors, the same assumption used by Badgwell et al. [5] and Coronell and Jensen [17]. More accurate temperature profiles in inlet and outlet zones would require adding thermocouples

in these zones.

2.4.1 Model verification

Fig. 2.2 shows the comparison of the model predicted wafer temperatures with the experimental data provided by Badgwell et al. [4]. The detailed reactor geometry is given in Table II. From Fig. 2.2 we observe that the model predicted temperature profile agrees with the measurement very well.

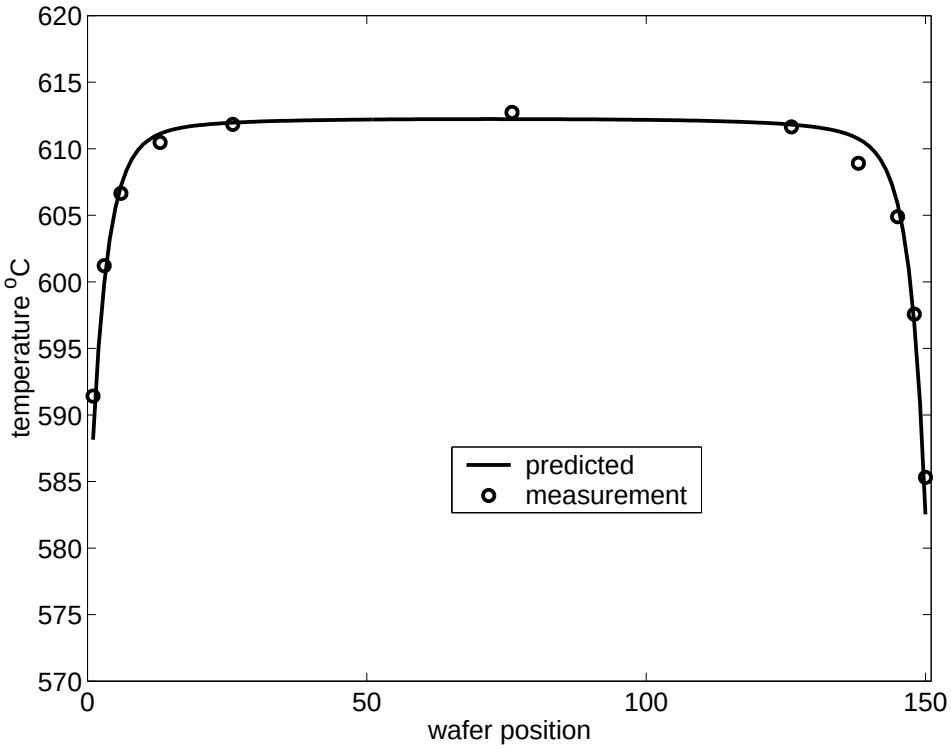


Figure 2.2: Comparison of the model predicted wafer temperatures with the experimental data [4]

2.4.2 Model sensitivity analysis

Since our model is transformed to a linear relationship, analytical sensitivity analysis can be straightforwardly performed, in contrast to other methods that require numerical solutions. Suppose we have solved the linear equation (2.23) for a particular set of base-case value of T_f, say T_f^*, and found a set of wafer temperatures $T_w = T_w^*$ that satisfied $(T_w^*)^4 = C(T_f^*)^4$. We want to consider the sensitivity of the T_w to T_f for perturbations around the base-case value T_f^*. Therefore, we rewrite (2.23):

$$\begin{bmatrix} T_{w1}^4 \\ T_{w2}^4 \\ \vdots \\ T_{wi}^4 \\ \vdots \\ T_{wm}^4 \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1j} & \cdots & C_{1n} \\ C_{21} & C_{22} & \cdots & C_{2j} & \cdots & C_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ C_{i1} & C_{i2} & \cdots & C_{ij} & \cdots & C_{in} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ C_{m1} & C_{m2} & \cdots & C_{mj} & \cdots & C_{mn} \end{bmatrix} \begin{bmatrix} T_{f1}^4 \\ T_{f2}^4 \\ \vdots \\ T_{fj}^4 \\ \vdots \\ T_{fn}^4 \end{bmatrix} \tag{2.24}$$

Taking the i^{th} row of the above matrix equation, we have:

$$T_{wi}^4 = C_{i1}T_{f1}^4 + C_{i2}T_{f2}^4 + \cdots + C_{ij}T_{fj}^4 + \cdots + C_{in}T_{fn}^4 \tag{2.25}$$

Taking the derivative on both sides of equation (2.25) with respect to T_{fj}, $j = 1, 2, \cdots, n$:

$$4\left(T_{wi}^*\right)^3 \left.\frac{\partial T_{wi}}{\partial T_{fj}}\right|_{T_w^*, T_f^*} = 4\left(T_{fj}^*\right)^3 C_{ij} \tag{2.26}$$

Solving equation (2.26) to get the sensitivity of T_{wi} to T_{fj} for perturbations around the base-case value T_f^* gives:

$$\left.\frac{\partial T_{wi}}{\partial T_{fj}}\right|_{T_w^*, T_f^*} = \left(\frac{T_{fj}^*}{T_{wi}^*}\right)^3 C_{ij} \tag{2.27}$$

Fig. 2.3 shows the sensitivity of the wafer temperatures to door temperatures where T_{fd} and T_{bd} denote the front and back door temperatures respectively. From the figure we see that the wafer temperatures are not sensitive to door temperatures, although end wafers are relatively much more sensitive than center wafers. Therefore, changing the door temperatures has little impact on the temperature uniformity across the wafer load.

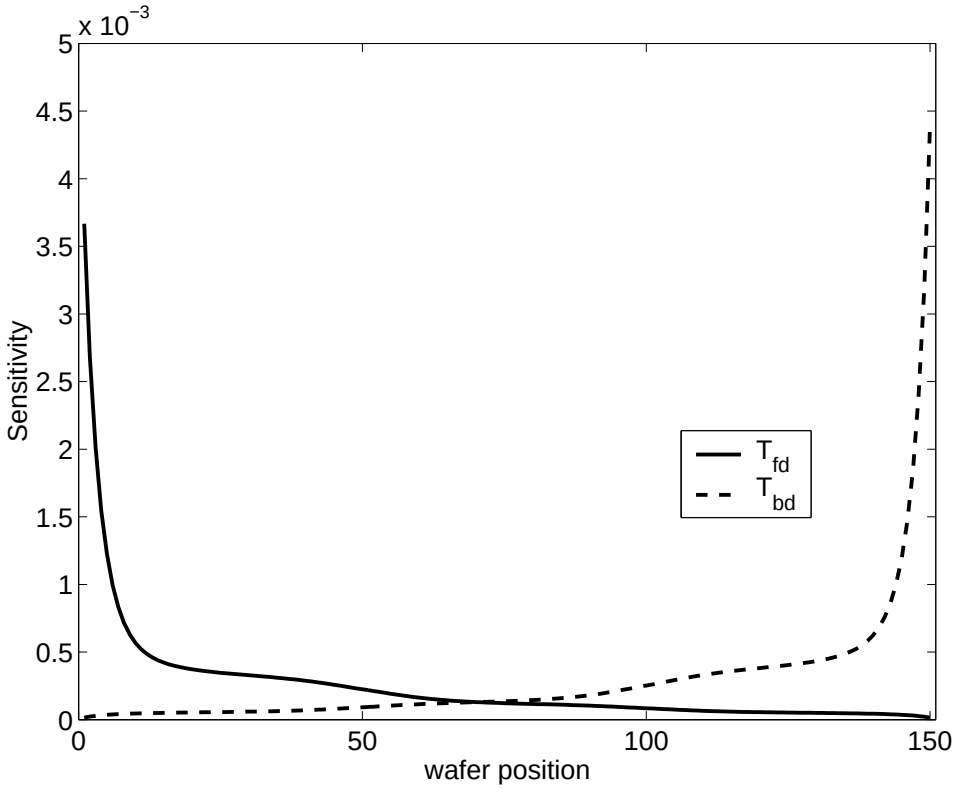


Figure 2.3: Sensitivity of wafer temperatures to door temperatures

Fig. 2.4 illustrates the sensitivity of the wafer temperatures to the temperatures of the five heating zone furnace elements, denoted by $T_{h1}, T_{h2}, \cdots, T_{h5}$. Although the first heating zone element is totally outside of the wafer load, wafers in the front of the load can still be affected. The second heating zone element covers about the first 1/3 of the wafer load. As we can see, when its temperature rises, the temperature of wafers in the next zone also rise. The third heating zone element covers the second 1/3 of the wafer load. Just like the second heating zone element, all wafer temperatures rise to different levels when the third zone temperature rises. The wafer temperature profile is coupled with the temperature of each heating zone element in complex ways. We cannot simply change some particular wafer temperatures without affecting others. Since the first and fifth elements have much more impact on the end wafers than on the center wafers in the wafer load, the temperature uniformity across the wafer load can be controlled by manipulating the temperatures of these two elements.

The sensitivity curves in Fig. 4 are very useful for selecting sensor locations for the furnace uniformity control. To achieve maximum sensitivity, sensors should be placed at the maxima of sensitivity curves. To ensure uniformity across all wafers, however, some sensors should be placed in between the maxima, where strong interaction between zones occurs. In practice the wafer temperatures cannot be reliably measured in-situ, but the wafer film thickness can be measured with post-process metrology, which can be used for run-to-run control. The control aspect of this process will be investigated in

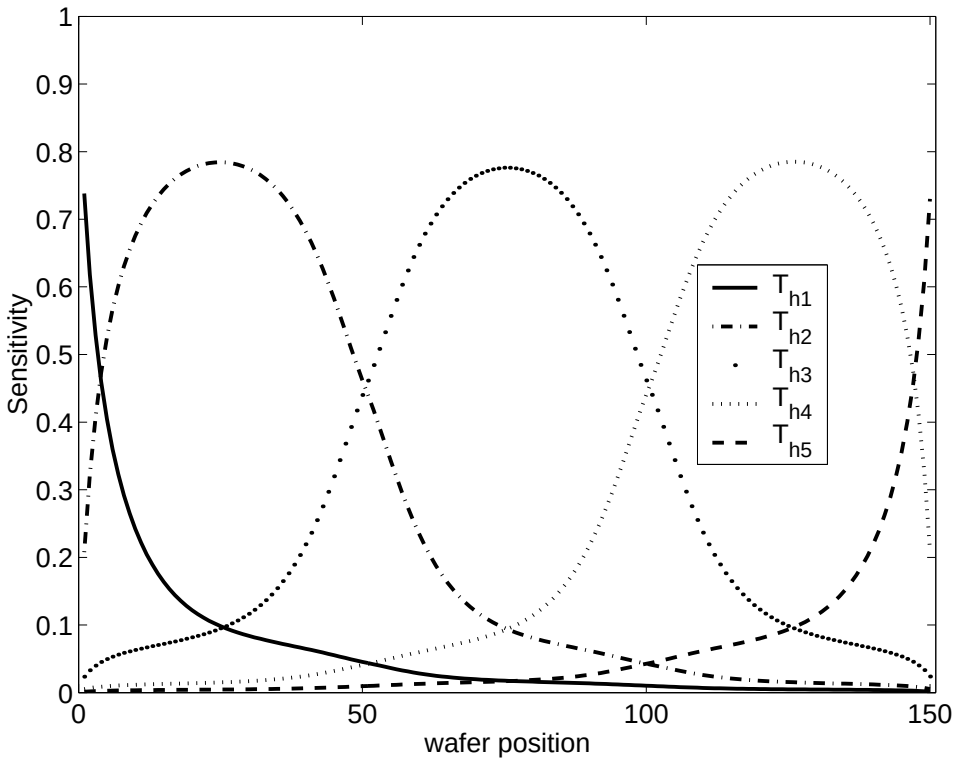


Figure 2.4: Sensitivity of wafer temperatures to temperatures of heating zone furnace elements

the future.

Fig. 2.5 illustrates the dependency of the wafer temperatures to the furnace wall emissivity where e_f denotes the furnace wall emissivity. Higher furnace wall emissivity will lead to higher wafer temperatures. The effect is approximately the same on all wafers in the wafer load. Fig. 2.6 shows that the door emissivity has almost no effect on the wafer temperature profile. Changes in the furnace emissivity do not appear to change the uniformity. Feedback control can be implemented to compensate for the effect on average thickness

of such emissivity changes.

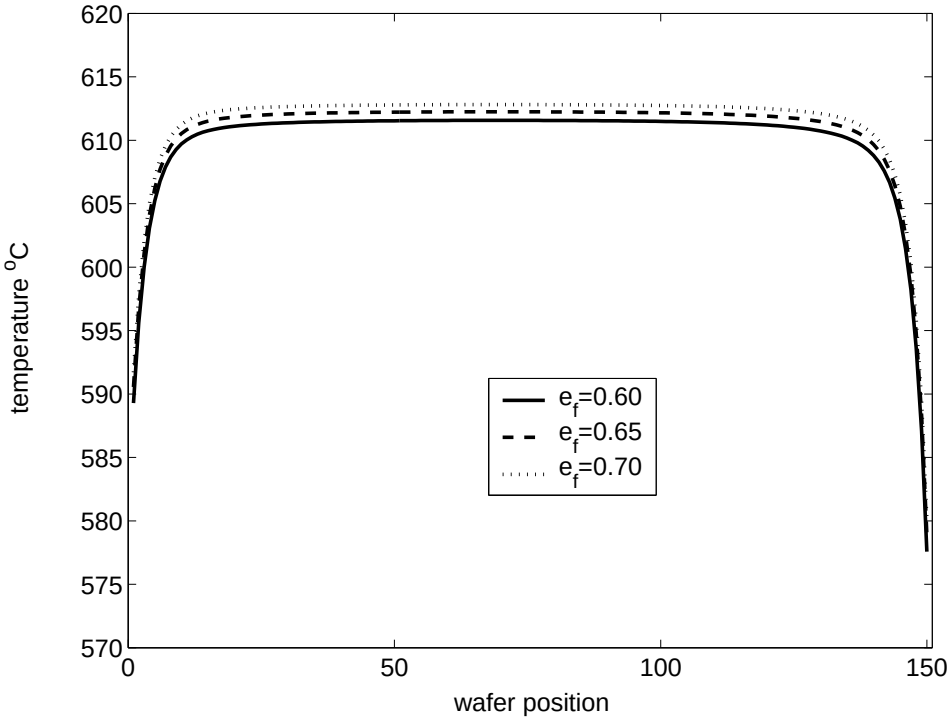


Figure 2.5: Wafer temperature dependence on furnace wall emissivity

2.5 Conclusions

In this work, the steady-state wafer temperature distribution in a hot-wall LPCVD reactor is modeled and solved analytically. A new first principles thermal model is developed to predict wafer temperatures from furnace wall temperatures based on an energy balance analysis. The predicted wafer temperatures show excellent agreement with published experimental data. The simple linear structure and light computation effort make this model useful

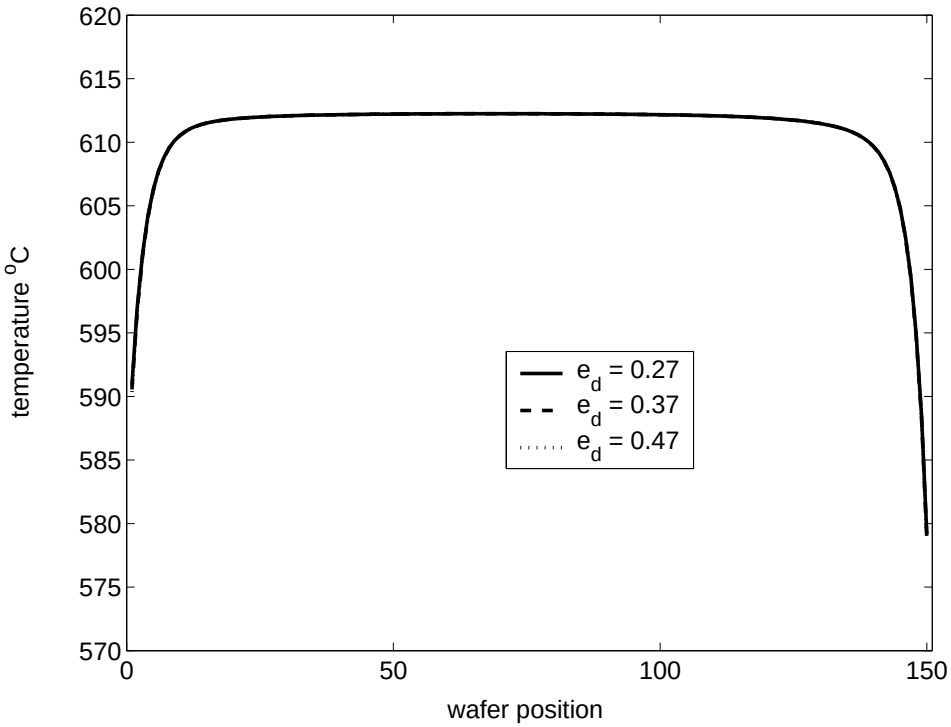


Figure 2.6: Wafer temperature dependence on door emissivity

for control in industry. Model sensitivity analyses are performed analytically. Sensitivity analyses show that wafer temperatures are sensitive to temperature changes of each furnace element in the heating zone, but not to those of the doors. This computationally efficient model can be used for real-time run-to-run control of across-load uniformity, which will be investigated in the future.

Chapter 3

Valve Stiction Modeling and Detection

Many control loops in process plants perform poorly due to valve stiction as one of the most common equipment problems. Valve stiction causes oscillation in control loops which increases variability in product quality, accelerates equipment wear, or leads to control system instability and other issues potentially disrupt the operation. Therefore, it is important to early detect valve stiction so that appropriate action can be taken to relieve the situation and avoid major shutdowns. Although there have been many attempts to understand and model valve stiction, those models are either physical models which are not practical to use, or empirical models but with complicated logic which makes them difficult to understand and implement. In this work, a new valve stiction model is proposed with a simple structure and straightforward logic which make it easy to implement. Furthermore, a new valve detection method is proposed based on curve-fitting of controller output signal for self-regulating processes or process output signal for integrating processes. A new metric referred to as a stiction index (SI), is introduced based on the proposed method to quantify the degree of valve stiction. Superior performance of the proposed method is demonstrated using both simulated data sets based on the new valve stiction model and real industrial data sets.

3.1 Introduction

Studies in the control performance monitoring show that in process industry many control loops perform poorly due to bad tuning or equipment problems [2, 24, 35, 81, 92], and it has been witnessed in some facilities that as high as one third of control loops are oscillating [20]. Oscillations in control loops raise particular concerns as they increase variability in product quality, accelerate equipment wear, and may cause other issues that could potentially disrupt the operation. Therefore, detecting and eliminating oscillations yield commercial benefits and are important activities in control loop supervision and maintenance.

In general, oscillations are caused by any one or a combination of the following reasons: (i) control valve stiction, (ii) poor controller tuning, (iii) poor process and control system design, and (iv) external oscillatory disturbances [8, 24, 74]. Simple and efficient methods have been developed to detect oscillating control loops automatically [26, 35, 74]. In this work, we focus on valve stiction detection given that oscillation has been detected.

To help understand the valve stiction phenomenon and simulate a sticky valve, several valve stiction models have been developed [14, 15, 64]. Choudhury et al. [14, 15] discuss the definition of stiction and distinguish it from other valve nonlinearities, and propose a data-driven model of stiction. Kano et al. [64] extend Choudhury's model to cope with both deterministic and stochastic signals. In this current work, the validity of these models is investigated and a new valve stiction model is proposed.

Several methods [19, 35, 54, 90, 97] have been developed to detect valve stiction in the last decade. However, all these methods require either detailed process knowledge or user interaction which are not desirable for automated monitoring systems [52]. Horch (1999) presented an automatic detection method based the cross-correlation function (CCF) between the controller output (OP) and the process output (PV) which is applicable to non-integrating processes. Later Horch [53] proposed another method to address the valve stiction in integrating processes by considering the probability distribution of the second derivative of controlled variable. In 2004, Singhal and Salsbury [89] proposed a valve stiction detection method based on the comparison of areas before and after the peak of an oscillating control error signal, *i.e.,* the difference between the set-point and the process variable being controlled. Kano et al. (2004) proposed two valve stiction detection methods, one requires knowing the valve position (VP) and the other is based on the plot of PV, OP with the shape of parallelogram. He and Pottmann (2003) developed a valve stiction detection technique [37] in which the OP is fitted piece-wisely to both triangular wave and sinusoidal wave using least squares method. A better fit to the triangle indicates valve stiction, while a better fit to the sinusoid indicates non-stiction. Also in that work, a stiction index (SI) was first defined as the ratio of the mean squared error (MSE) of sinusoidal fitting and the sum of the MSE's of both sinusoidal and triangular fittings. An SI close to zero would indicate non-stiction while an SI close to one would indicate stiction. In the meantime, Rossi and Scali (2004) proposed a very

similar technique independently, in which the PV instead of OP is fitted using three different models: relay wave, triangular wave and sinusoidal wave [85].

In this work, we extend our 2003 work to cover both self-regulating and integrating processes based on the following observations: In the case of control loop oscillation caused by poor controller tuning or external oscillating disturbance, the OP and PV typically follow sinusoidal waves for both self-regulating and integrating processes. In the case of stiction, for self-regulating processes, the OP will move like a triangular wave, while for integrating processes such as level control, the PV will move like a triangular wave. The basic idea of the new detection method is to fit two different functions ,triangle and sinusoid, to the measured oscillating signal (OP for self-regulating processes and PV for integrating processes). A better fit to the triangle indicates valve stiction, while a better fit to the sinusoid indicates non-stiction. The SI metric is used as a criterion to evaluate the existence of valve stiction.

The remainder of this paper is organized as follows: Section 2 reviews valve stiction models and presents a new valve stiction model with simple structure and straightforward logic. Section 3 reviews published stiction detection techniques and analyzes Horch's first method in detail. Also in Section 3, a new valve stiction detection method and its theoretical analysis are presented, together with the simulation demonstrations. The application of the proposed method to industrial examples is presented in Section 4. Some conclusions are drawn in Section 5.

3.2 Valve Stiction Model

The purpose of this section is to understand the characteristics of valve stiction and mathematically reproduce its behavior. Literally, valve stiction can be represented as the necessary force applied to the valve stem to make it move [30]. Due to stiction, the valve will not move if the amount of force corresponding to the controller output is too small to overcome the static friction. Because the controller output (OP) adjustment is not materialized by the actuator (*i.e.,* the valve) due to stiction, integral action in the controller will cause the OP continue to increase in the same direction until the valve overcomes the stiction band. Once overcome, the valve moves suddenly with more than the desired amount causing the process to overreact. The OP then changes in the opposite direction trying to get the process back on track until the valve overcomes the stiction band, which makes the process overreact again in the opposite direction, thus causing process oscillation.

Two types of models have been developed to simulate valve stiction. One is detailed physical models [15] that formulate the stiction phenomenon using the force balance based on Newton's second law of motion. The other is empirical data-driven models [14, 15, 64] that describe the relationship between the OP and the valve position or valve output (VP). The main disadvantage of a physics-based model is that it requires knowledge of several parameters such as the mass of the moving part and different types of friction forces which cannot be easily measured and change with the type of flowing fluid and part wearing. Therefore, in this work, we focus on the data-driven model, and

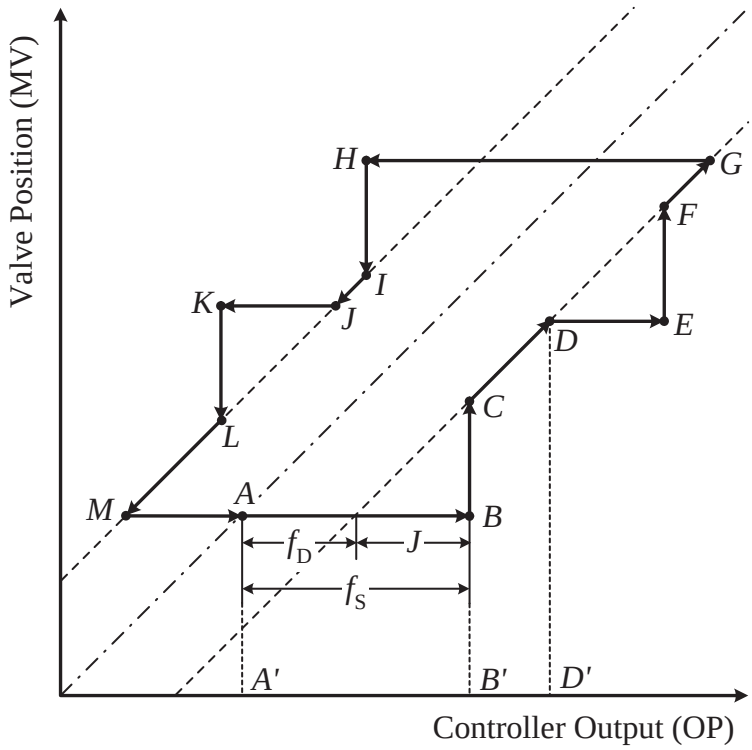


Figure 3.1: Schematic operation diagram of a sticky valve

models developed by Kano et al. (2004) and Choudhury et al. (2004, 2005) are discussed here.

Figure 3.1 shows the typical input-output behavior of a sticky valve. Without stiction, the valve would move along the dash-dot line crossing the origin, *i.e.,* any amount of OP adjustment would result in the same amount of VP change. However, for a sticky valve, static and kinetic frictions have to be taken into account. In the figure, f_S, f_D and J denote static friction band, kinetic friction band and stick band respectively. Because stiction is

generally measured as percentage of the valve travel range, for simplicity, as in Choudhury et al. (2004) and Kano et al. (2004), all variables such as f_S, f_D, J, controller output u [1], process output y and valve position u_V are translated to percentage of valve range so that algebra can be performed among them directly. For example, J is defined as:

$$J = f_S - f_D \tag{3.1}$$

To illustrate how OP adjustment drives VP change in a sticky valve in Figure 3.1, suppose the valve rests at a neutral position A at the beginning. If the OP adjustment is between $A'B'$, the valve will not be able to overcome the static friction band f_S so the VP will not change. However, if the OP moves outside of $A'B'$, say D', then the valve is able to overcome f_S at point B and jumps to point C. After that, the valve moves from C to D, overcoming f_D only.

In Kano's model, they also define the summation of static and kinetic friction bands S:

$$S = f_S + f_D \tag{3.2}$$

The flow chart of Kano's stiction model is shown in Figure 3.2 which is very similar to Choudhury et. al (2004). Some notations are explained below: Two states of the valve are distinguished by *stp*: *stp* is reset to 0 if the controller output $u(t)$ results in a valve move, otherwise is reset to 1. $d = \pm 1$ denotes

[1] In this work, both OP and u stand for controller output. OP is usually used in descriptions, tables and figures; u is usually used in mathematical derivations. Similarly, both PV and y denote process variable; both VP and u_V denote valve position or valve output.

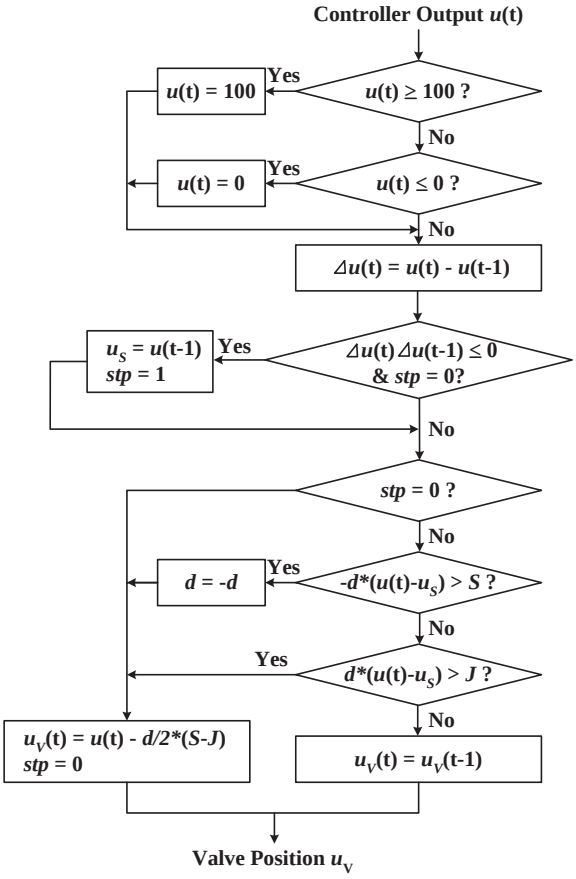


Figure 3.2: Flowchart of Kano's valve stiction model [64]

the direction of frictional force. The model structure is complicated and it is difficult to get the logic straight. However, if we examine it carefully, there is a deficiency in this model. In the case where the previous controller output $u(t-1)$ resulted in a valve move so that $stp = 0$, and the adjustments made on controller output for this run and previous run are in the same direction so that $\Delta u(t)\Delta u(t-1) > 0$, no matter how small $\Delta u(t)$ is, according to the

model, the VP will always change. To explain it graphically, we assume $u(t-1)$ resulted the valve movement along $A \rightarrow B \rightarrow C \rightarrow D$ and finally stuck at D as shown in Figure 3.1 (assuming $\Delta u(t-1) > 0$ and stp $= 0$). In the next run, if $\Delta u(t) > 0$ ($u(t)$ moves along the positive direction), logically, the VP should move along $D \rightarrow E \rightarrow F \rightarrow G$. If the adjustment made is too small to overcome the static friction ($0 < \Delta u(t) < J$), the VP should not change (stuck somewhere between D and E). But according to Kano's model, no matter how small $\Delta u(t)$ is, as long as it is greater than zero, the valve will always move and stop somewhere between D and F (dash line) which is not logically correct. Choudhury's model has the same problem.

Another drawback associated with Kano and Choudhury's models is that the saturation constraints are added to the controller output instead of actuator (valve). Based on the typical input-output behavior of a sticky valve, we propose a new valve stiction model. Figure 3.3 shows the flowchart of the new model, which is much simpler and more straightforward in logic. If desired, the saturation constraint can be easily added to $u_V(t)$ after the model calculation.

3.3 Valve Stiction Detection

In this section, we briefly review several published detection methods first, then we examine Horch's first method [52] in detail and we show that it is not consistent.

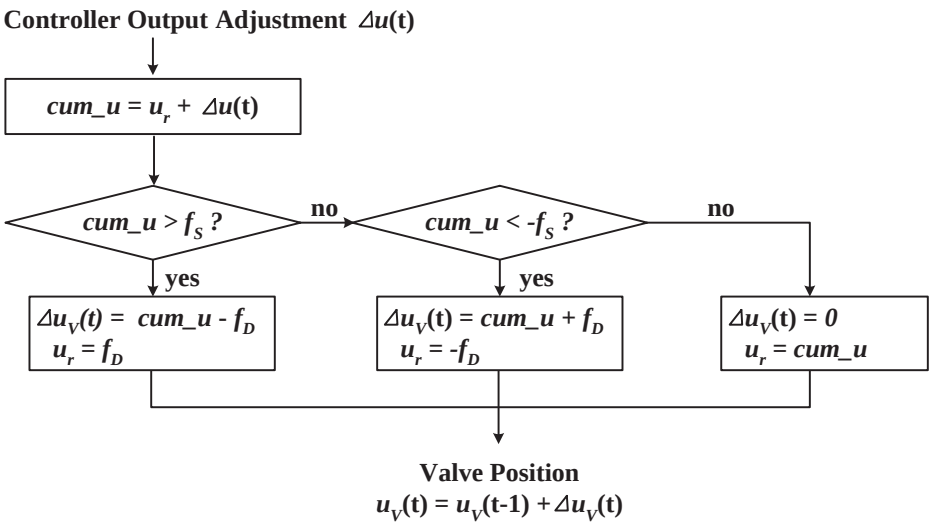


Figure 3.3: Flowchart of proposed valve stiction model

3.3.1 Published valve stiction methods

In 1999, Horch proposed a valve stiction detection method for self-regulating processes based on the CCF between OP and PV. Given the following assumptions: (i) the process does not have an integral action, (ii) the process is controlled by a PI controller, and (iii) the process is oscillating with a significantly large amplitude, Horch claims that the valve stiction would result in an odd CCF, *i.e.,* the phase shift between OP and PV is $\pi/2$, while an external oscillating disturbance or an aggressive controller would result in an even CCF, *i.e.,* the phase shift between OP and PV is π. The method is proposed based on observations and verified by some simulated and industrial data and its validity is investigated in Section 3.3.2.

Later Horch [53] proposed another method to address the valve stiction in integrating processes by considering the probability distribution of the second derivative of controlled variable – in the stiction case the distribution is close to Gaussian, otherwise it will have two peaks. One drawback of this method is the differentiation of noisy signals. A suitable filter and cut-off frequency have to be carefully chosen in order to filter out noise. This can hardly be done automatically since different processes have different system characteristics and different noise levels. It has been observed that even after filtering, the calculation of derivatives amplified moderate amount of noise and blurred the distinction between the shapes of the two probability distributions [89].

In 2004, Singhal and Salsbury proposed a valve stiction detection method based on the comparison of areas before and after the peak of an oscillating control error signal (*i.e.,* the difference between the set-point and the process variable being controlled). The idea is based on the observation that aggressive controller usually results in a sinusoidal control error signal, while for a sticking valve, the oscillating signal typically rises slower than drops. There are several practical considerations as mentioned by authors: (i) the methodology can not be applied to integrating processes, (ii) the methodology cannot distinguish other nonlinearities from stiction, (iii) the error signal must be sampled many times per oscillation period in order to get accurate peak location and areas calculation, and (iv) noise adds variation to the peak and zero-crossing locations which can result in misleading diagnosis.

Kano et al. (2004) proposed two valve stiction detection methods.

Method A is based on the percentage of time when VP does not change while OP changes. Method B is based on the plot of PV vs OP takes shape of parallelogram. However, as pointed out by the authors, these methods should be used only when flow rate or valve position is measured. Method B is not always reliable even when flow rate or valve position is measured as shown in one of their flow control examples.

Rossi and Scali (2004) proposed a technique to fit the PV using three different models: relay wave, triangular wave and sinusoidal wave. Relay and triangular waves are associated with the presence of stiction, while sinusoidal shape with the presence of external perturbations [85]. Although this method is very similar to the method we propose in this work, it looks at PV only and claimed by authors that it is applicable to self-regulating processes only.

3.3.2 Analysis on Horch's first method

It has been discussed in others' work [85, 89] that Horch's first method does not work very well all the time. However, no theoretical analysis has been given to prove that Horch's first method is inconsistent. In this subsection, we show that with no valve stiction, different controller tuning could result in either an odd CCF or an even CCF between OP and PV, and thus demonstrates the inconsistency of the method.

It is a common practice to approximate a higher order system using a first order plus time delay model (FOPTD). Without loss of generality, we consider the plant as an FOPTD process, and the controller is a PI controller,

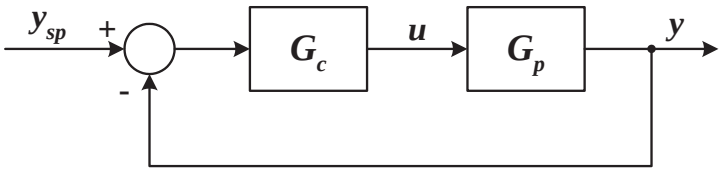


Figure 3.4: Block diagram of an FOPTD process under PI control

as shown in Figure 3.4. The analysis can be easily extended to higher order systems.

The process transfer function is given as

$$G_p = \frac{K_p \cdot e^{-\theta s}}{\tau s + 1} \tag{3.3}$$

where K_p is the process gain, θ is the process delay and τ is the process time constant. The dynamics of the control valve can be approximated as a first order system. However, the time constant of a control valve is usually much smaller than the process time constant, for simplicity, its dynamics is ignored here.

The controller transfer function is

$$G_c = K_c(1 + \frac{1}{\tau_I s}) \tag{3.4}$$

where K_c is the proportional gain and τ_I is the integral time constant for the PI controller.

Now we show that with no valve stiction, different controller tuning could result in either odd or even CCF between OP and PV. It is straightforward to show that the phase shift between OP and PV is different at different

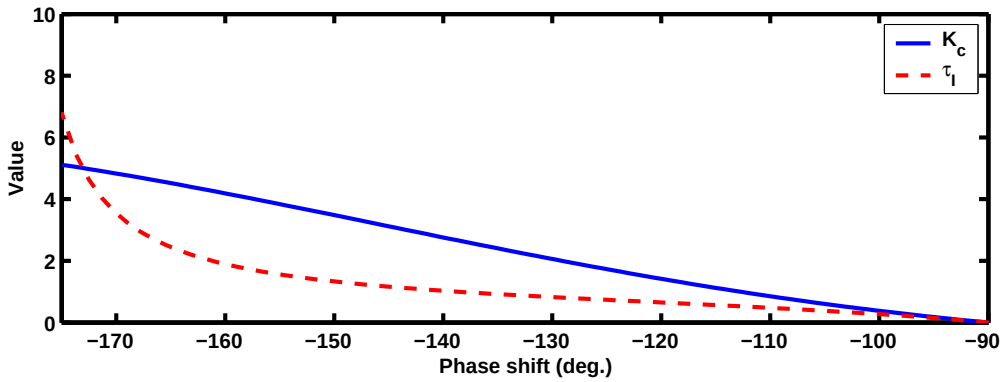


Figure 3.5: The relationship between (K_c, τ_I) and phase shift at marginal stable condition

frequencies. Therefore, controller tuning can result in either odd or even CCF between OP and PV.

The oscillation frequency can be obtained by solving the following characteristic equation,

$$1 + G_c(j\omega)G_p(j\omega) = 0 \tag{3.5}$$

Plugging Equation (3.3) and (3.4) into Equation (3.5) and applying Euler's formula we have

$$-\tau_I\tau\omega^2 + j\tau_I\omega + K_cK_p(cos\omega\theta - jsin\omega\theta)(1 + j\tau_I\omega) = 0 \tag{3.6}$$

It is straightforward to see that the phase shift ϕ at certain frequency ω is,

$$\phi = \angle\, G_p(j\omega) = \frac{\pi}{2} + \alpha - \omega\theta \tag{3.7}$$

where $\alpha \equiv \arctan(1/\omega\tau)$. Equation (3.7) shows that under different controller tuning, it will result in different phase shift between OP and PV. To illustrate

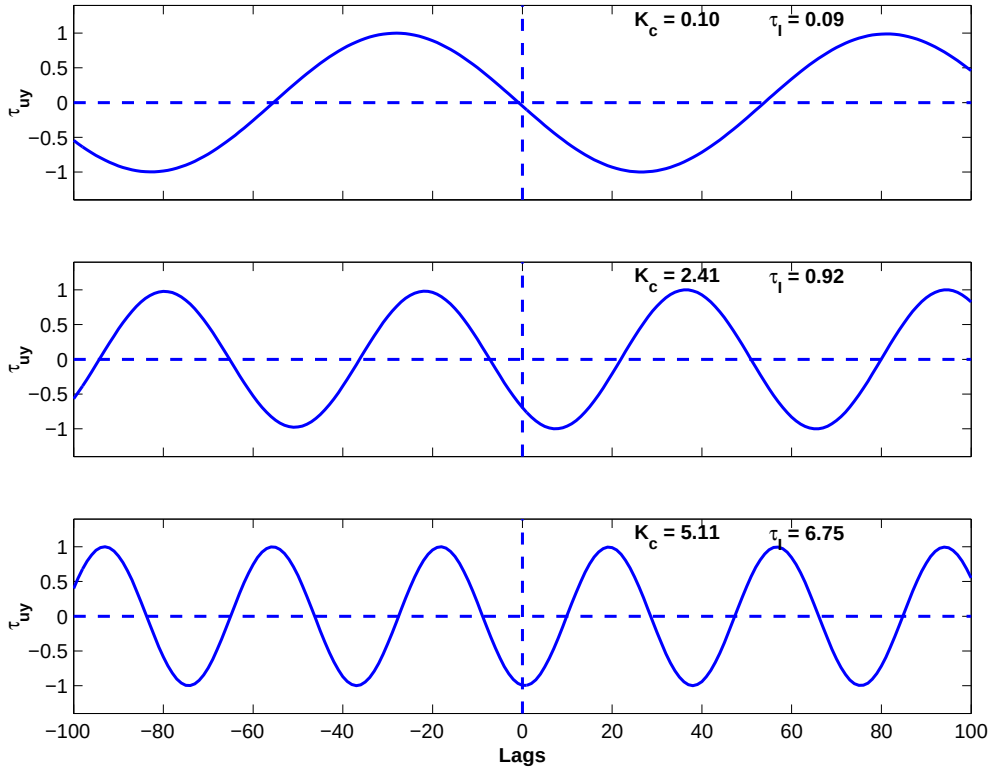


Figure 3.6: Different controller tunings result in different types of CCF between OP and PV

this point, we look at the following example with the plant model:

$$G_p = \frac{e^{-s}}{3s+1} \tag{3.8}$$

The relationship between phase shift ϕ and controller tuning (K_c & τ_I) is shown in Figure 3.5 where different tuning can result in different phase shift ranging from $-\pi$ to $-\pi/2$. More specifically, we pick three pairs of K_c & τ_I at phase shift about $-\pi$, $-\frac{3\pi}{4}$ and $-\pi/2$ and the resulted CCF's are shown in

Figure 3.6. Horch's first method would conclude that there is no stiction for the first case, undetermined for the second case and stiction for the third case, although the truth is that there is no stiction but different controller tunings.

3.4 Proposed Curve Fitting Method

In this subsection, a simple curve fitting method is proposed for valve stiction detection based on the original work we did in 2003 given the following assumptions:

(i) The control loop has been detected as oscillating.

(ii) There are enough points sampled during each oscillation period.

The first assumption makes sense because any valve stiction, if significant enough to be a concern, would cause loop oscillating, and it is not necessary to check valve stiction if the control loop is not oscillating. The second assumption is necessary in order to precisely distinguish two different types of signals - sinusoidal and triangular waves. Although the more sample points we have, the more reliable the results will be, our experience shows that 7 to 8 sample points per half oscillation period would be sufficient.

The key idea of proposed method is to make use of the following observations:

(i) In the case of stiction in a control valve, the valve position switches back and forth intermittently, which results in a rectangular wave signal. The integrator in the PI controller (or in the process if it is an integrating process)

integrates the rectangular wave into triangular wave.

(ii) An oscillatory external disturbance usually results in sinusoidal controller output and process output signals.

(iii) A marginal stable control loop also results in smooth sinusoidal shape controller output and process output.

The main idea is to fit two different functions ,triangle and sinusoid , to the measured controller output for self-regulating processes or process output for integrating processes. A better fit to the triangle indicates valve stiction, a better fit to the sinusoid, non-stiction. A new metric – Stiction Index (SI) is defined and used as a criterion to estimate the probability of valve stiction.

3.4.0.1 Theoretical derivation

The proposed stiction detection method is designed for both self-regulating and integrating processes under PI control. For a self-regulating process, the plant is approximated by an FOPTD model as in Section 3.3.2 and the following derivation for self-regulating processes is based on the model given by Equation (3.3).

While for an integrating process, *e.g.,* level control, we use the following simple model

$$G_p = \frac{K_p e^{-\theta s}}{s} \tag{3.9}$$

A PI controller is used to control the plant and its transfer function is given in Equation (3.4). The overall system we consider is shown in Figure 3.4. Note

that the analysis can be extended to the higher order system straightforwardly.

From the previous discussion we know that when there is a valve stiction, it results in valve output (VP) that consists of a sequence of rectangular pulse signals. Therefore, the process output (PV) is a sequence of plant step responses. To simplify the derivation, let the step size be one. Because the effect of time delay θ in Equation (3.3) is just a time shift of the occurrence of the step response, here we consider the case where $\theta = 0$ and we examine the self-regulating processes and integrating processes separately.

Self-regulating processes

For self-regulating processes, we examine the shape of controller output $u(t)$. The process step response is given by:

$$y(t) = y_0 + K_p(1 - e^{-\frac{t}{\tau}}) \tag{3.10}$$

where y_0 is the initial value of PV when VP switches. The deviation of the output from the target, *i.e.,* input to the PI controller, is:

$$e = y_{sp} - y(t) = y_{sp} - y_0 - K_p(1 - e^{-\frac{t}{\tau}}) \tag{3.11}$$

and the controller output (OP) is

$$\begin{aligned} u(t) &= K_c\left[y_{sp} - y_0 - K_p(1 - e^{-\frac{t}{\tau}})\right] + \frac{K_c}{\tau_I}\left[y_{sp} - y_0 - K_p\right] \cdot t + K_c K_p \int_0^{t_f} e^{-\frac{t}{\tau}} dt \\ &= K_c\left[y_{sp} - y_0 - K_p\right] + K_c K_p \tau[1 - e^{-\frac{t_f}{\tau}}] + \frac{K_c}{\tau_I}\left[y_{sp} - y_0 - K_p\right] \cdot t \\ &\quad + K_c K_p e^{-\frac{t}{\tau}} \end{aligned} \tag{3.12}$$

Equation (3.12) shows that the first three terms correspond to a straight line with slope $\frac{K_c}{\tau_I}[y_{sp} - y_0 - K_p]$, and the last term corresponds to an exponential decay on top of the straight line. When the process has a fast dynamic, *i.e.,* a small τ, then $K_c K_p e^{-\frac{t}{\tau}} \doteq 0$, which makes VP follow a straight line with slope $\frac{K_c}{\tau_I}[y_{sp} - y_0 - K_p]$; while if the process has a slow dynamic, *i.e.,* a large τ, then $K_c K_p e^{-\frac{t}{\tau}} \doteq -\frac{K_c K_p}{\tau} t$, which makes VP follow a straight line with slope $\frac{K_c}{\tau_I}[y_{sp} - y_0 - K_p] - \frac{K_c K_p}{\tau}$. Even for the case where the exponential term is significant as shown in Figure 3.7 (a), since we piece-wisely fit half period of oscillating signal to sinusoid or triangle, as we can see from Figures 3.7 (b) and (c), triangle is strongly favored over sinusoid [2]. To conclude, for stiction in self-regulating processes, OP can be approximated by triangular wave.

Integrating processes

For integrating processes, we examine the shape of process variable $y(t)$. Given the process transfer function in Equation (3.9), the step response is

$$y(t) = y_0 + K_p t \tag{3.13}$$

Equation (3.13) shows that the step response of an integrating process is a straight line so that PV can be approximated by triangular wave.

[2] Sinusoid and triangle fittings are discussed in Section 3.4.0.2.

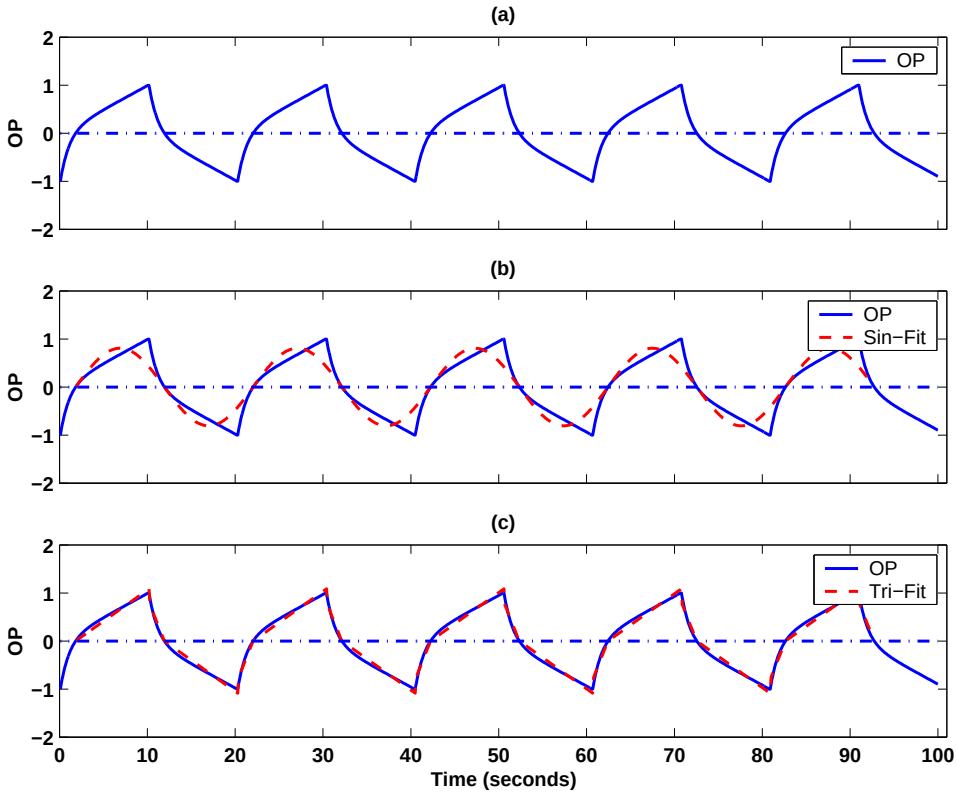


Figure 3.7: Curve fitting of OP of a self-regulating process in case of stiction

3.4.0.2 Curve fitting

It is assumed that the loop in question is known to be oscillating, *e.g.,* by using methods proposed in [26, 35, 74]. After the detection of the oscillation, the signal is detrended and mean-centered. The location of each zero-crossing is automatically detected, and determined by linear interpolation of two points on both sides of axis.

Sinusoid fitting

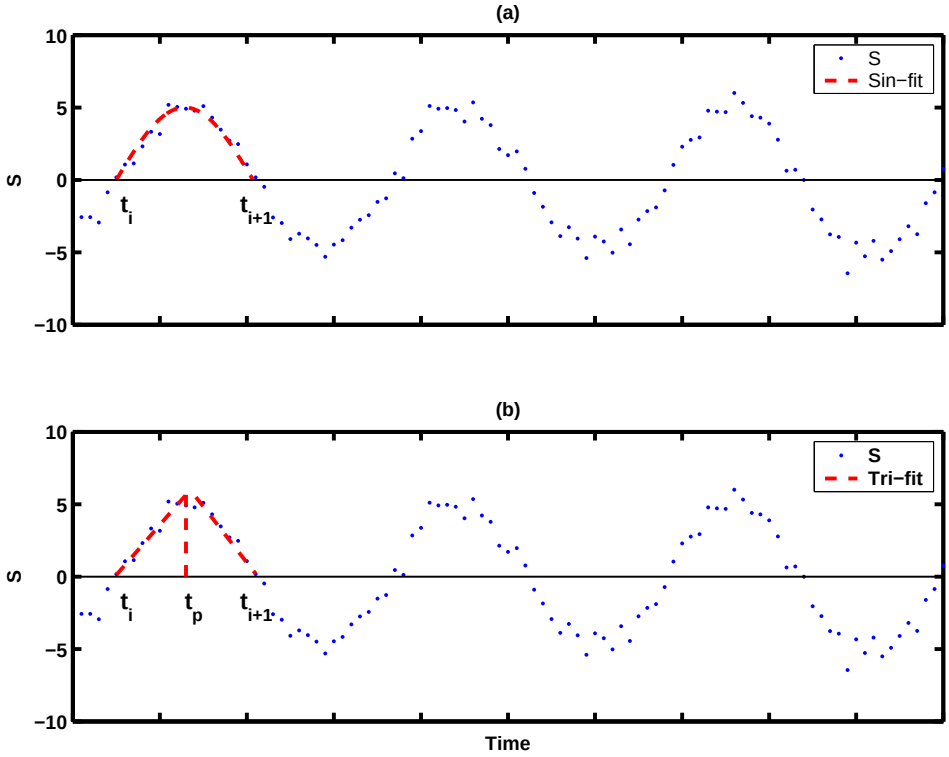


Figure 3.8: Curve fitting: (a) sinusoid fitting; (b) triangle fitting

The curve is fitted piece-wisely for each half-period of oscillation (see Figure 3.8 (a)), which means, each fitting piece may have different amplitude and/or frequency. This consideration is reasonable considering the presence of noise in the signal. Besides, in real process, the oscillation magnitude may change from time to time and other factors (*e.g.*, external disturbances) may result in an unsymmetrical signal with respect to its mean.

Denoting the signal to be fitted as S, for best sinusoid fitting of each

half-period, our objective function is:

$$J = min\left\{xsin\left(\omega(t_i : t_{i+1} - t_i) + \phi\right) - S(t_i : t_{i+1})\right\} \tag{3.14}$$

where x is the amplitude, ω the frequency and ϕ the phase shift of the sinusoid. $(t_i : t_{i+1})$ is the time range of fitting as in Figure 3.8 (a). In our case, because the curve is fitted piece-wisely, we have $\phi = 0$. For simplicity, we fix ω to be

$$\omega = \frac{2\pi}{t_i - t_{i+1}} \tag{3.15}$$

So our optimization problem is to find x which minimize the difference between the fitted curve and the signal S. By defining two vectors

$$a \equiv sin\left(\omega(t_i : t_{i+1} - t_i) + \phi\right) \tag{3.16}$$

$$b \equiv S(t_i : t_{i+1}) \tag{3.17}$$

we have

$$J = min\left\{ax - b\right\} \tag{3.18}$$

By using simple least squares method, we have

$$x = (a^T a)^{-1}(a^T b) \tag{3.19}$$

After the optimal x is determined, the mean squared error $\text{MSE}_{Sin}(i)$ for the sinusoid fitting during time period $(t_i : t_{i+1})$ is calculated. The overall mean squared error for sinusoid fitting MSE_{Sin} is the average of $\text{MSE}_{Sin}(i)$ over all time periods.

Triangle fitting

Triangle fitting is more difficult because it is a piece-wise curve fitting with two degrees of freedom: the location and the magnitude of the maxima. So we use numerical iterative method to find the best fitting. The algorithm is described below:

Step 1: For each half-period of signal S (*e.g.,* t_i to t_{i+1}), set the minimum MSE: $\text{MSE}_{Tri}(i) = \infty$.

Step 2: For peak location t_p from t_i to t_{i+1}, find the first linear least squares fitting for $(t_i : t_p)$ with constraint that the line has to pass the first zero-crossing point at t_i and the second linear least squares fitting for $(t_{p+1} : t_{i+1})$ with constraint that the line has to pass the second zero-crossing point at t_{i+1}. Then calculate MSE between t_i and t_{i+1}. If $\text{MSE}_{Tri}(i) >$ MSE, set $\text{MSE}_{Tri}(i)$ = MSE.

Step 3: Repeat step 2 for different t_p's.

Step 4: Repeat steps 1, 2 and 3 for each piece of S: $i = i + 1$.

Step 5: The overall MSE_{Tri} is the average of the minimum MSE of each piece – $MSE_{Tri}(i)$.

One example of triangle fitting is given in Figure 3.8 (b).

3.4.0.3 Stiction index (SI)

SI is defined as the ratio of the mean squared error of the sinusoid fitting to the summation of the mean squared errors of both sinusoid and

triangle fittings:

$$SI = \frac{MSE_{Sin}}{MSE_{Sin} + MSE_{Tri}} \tag{3.20}$$

Note that SI is bounded to the interval [0, 1]. $SI = 0$ indicates non-stiction where S fits sinusoid perfectly ($\text{MSE}_{Sin} = 0$), while $SI = 1$ indicates stiction where S fits triangle perfectly ($\text{MSE}_{Tri} = 0$). The real process data will not show such ideal clear-cut separation, an SI close to 0 would indicate non-stiction while an SI close to 1 would indicate stiction. We would recommend the following rules:

$$\begin{aligned} \text{SI} &\leq 0.4 \implies \text{No stiction} \\ 0.4 < \text{SI} &< 0.6 \implies \text{Undetermined} \\ \text{SI} &\geq 0.6 \implies \text{Stiction} \end{aligned}$$

3.5 Simulation Examples

In this subsection, the proposed valve stiction detection method is applied to simulation examples in which the stiction is introduced using the proposed stiction model. To compare our valve detection method with Kano's methods, the same flow control and level control systems used in Kano et. al (2004) are investigated. Block diagrams of two systems are shown in Figures 3.9 and 3.10 and process transfer functions are given by:

$$G_f(s) = \frac{1}{0.2s + 1} \tag{3.21}$$

$$G_l(s) = \frac{1}{15s} e^{-s} \tag{3.22}$$

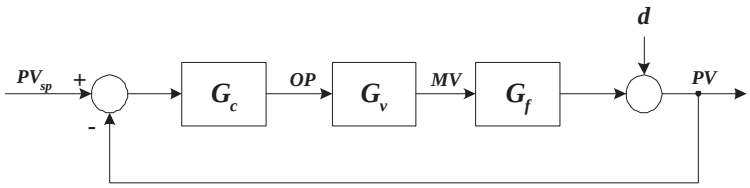


Figure 3.9: Block diagram of flow control system

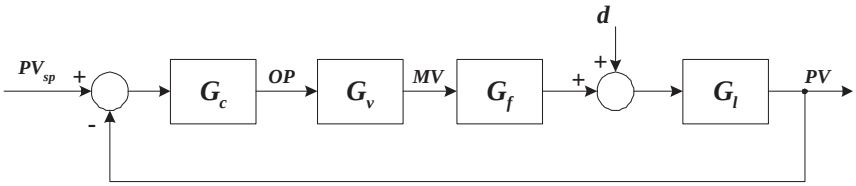


Figure 3.10: Block diagram of level control system

Table 3.1: Valve stiction model parameters [64]

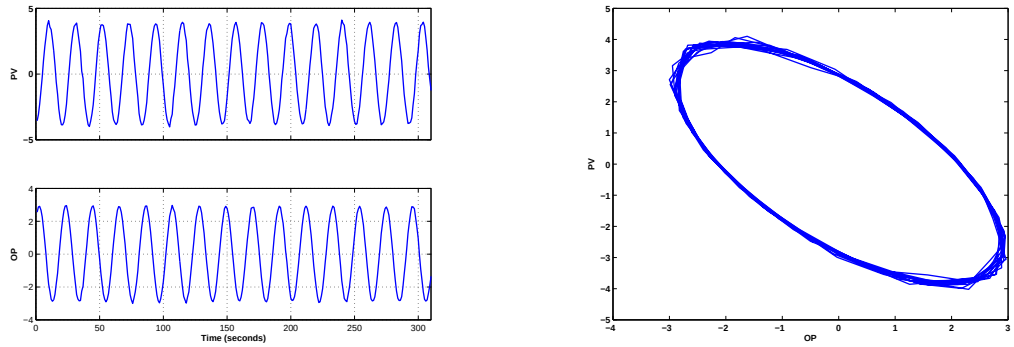
Case number	Degree of stiction	f_D	f_S
Case 1	No stiction	0	0
Case 2	Weak stiction	0.35	0.65
Case 3	Strong stiction	2	3

PI controllers are used for both control systems and their transfer functions are given by:

$$G_{c_1} = 0.5\left(1 + \frac{1}{0.3s}\right) \tag{3.23}$$

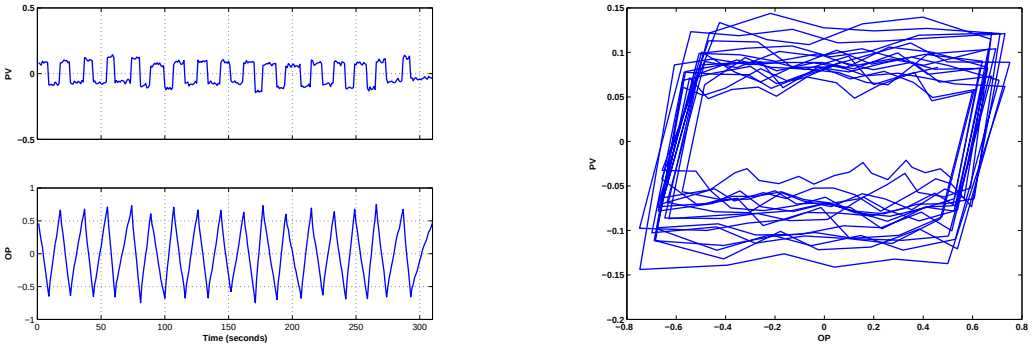
$$G_{c_2} = 3\left(1 + \frac{1}{30s}\right) \tag{3.24}$$

Three cases are examined for both systems: no stiction, weak stiction and strong stiction. Valve stiction model parameters are summarized in Table 3.1.



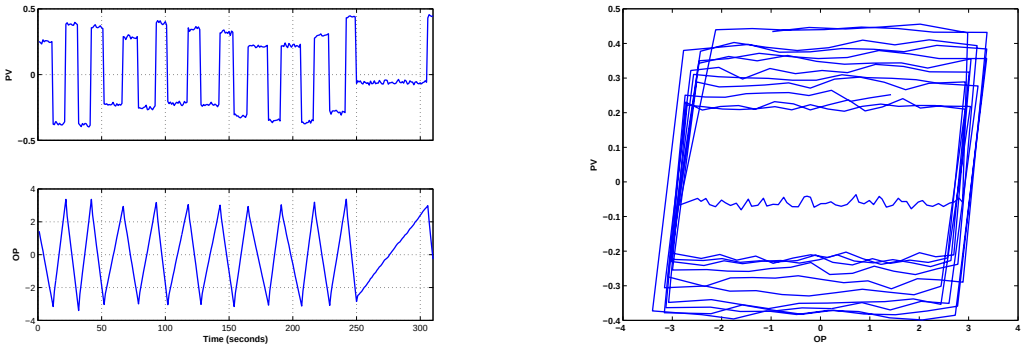
(a) (b)

Figure 3.11: Flow control, case 1 – no stiction, but external sinusoidal disturbance



(a) (b)

Figure 3.12: Flow control, case 2 – weak stiction



(a) (b)

Figure 3.13: Flow control, case 3 – strong stiction

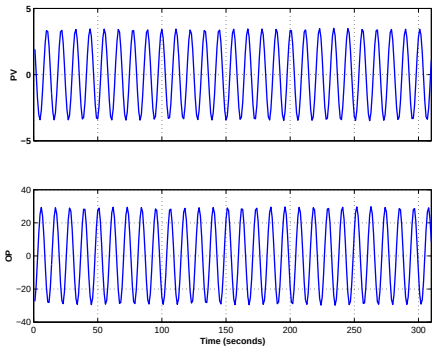
Table 3.2: Flow control case study

Control system	Case number	SI_{OP}
Flow control	Case 1	0.04
Flow control	Case 2	0.98
Flow control	Case 3	1.00

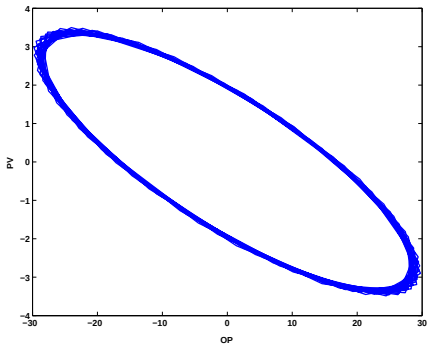
In the case of no stiction, because both systems are closed-loop stable systems, there is no oscillation and the proposed method is not applicable. To test the capability of the proposed method on distinguishing valve stiction from external disturbance, for the case of no stiction in flow control system, an external sinusoidal disturbance is introduced:

$$d = 5sin(3t) \tag{3.25}$$

Simulation results for flow control are shown in Figures 3.11, 3.12 and 3.13 and detection results are listed in Table 3.2. As we can see, the stiction index based on controller output (SI_{OP}) successfully detected valve stiction in flow control and distinguished it from an external disturbance.

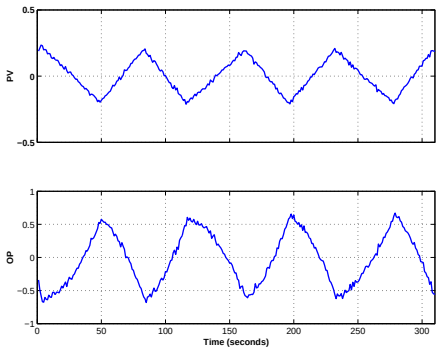


(a)

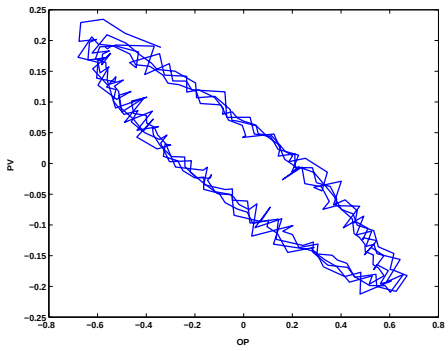


(b)

Figure 3.14: Level control, case 1 – no stiction, but aggressive tuning

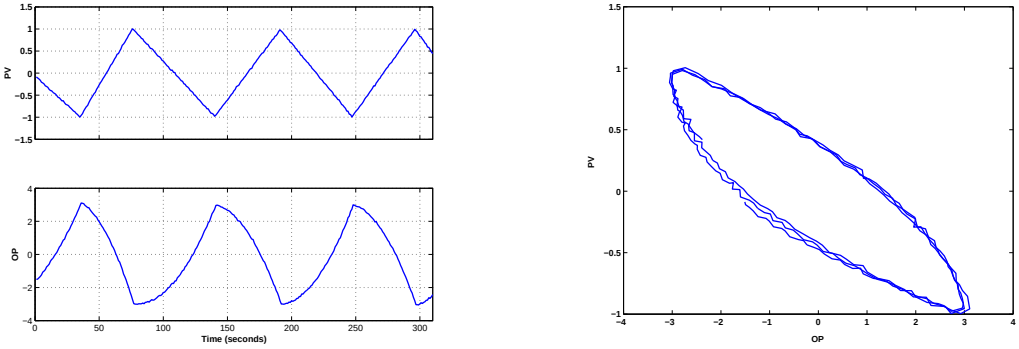


(a)



(b)

Figure 3.15: Level control, case 2 – weak stiction



(a) (b)

Figure 3.16: Level control, case 3 – strong stiction

Table 3.3: Level control case study

Control system	Case number	SI_{PV}
Level control	Case 1	0.00
Level control	Case 2	0.80
Level control	Case 3	0.99

To test the capability of distinguishing valve stiction from bad tuning, for the case of no stiction in level control system, the controller gain is increased from 3 to 8.4 to make the system marginal stable. Simulation results for level control are shown in Figures 3.14, 3.15 and 3.16 and detection results are listed in Table 3.3. The stiction index based on process output (SI_{PV}) detected stiction successfully and distinguished it from bad tuning.

As a comparison, in the case of flow control, Kano's method A [64] detects the stiction successfully, but Kano's method B fails. In the case of level control where level is used for detection, none of Kano's methods can detect stiction successfully. For the cases where there might be multiple causes

Table 3.4: Application results with mixed cases

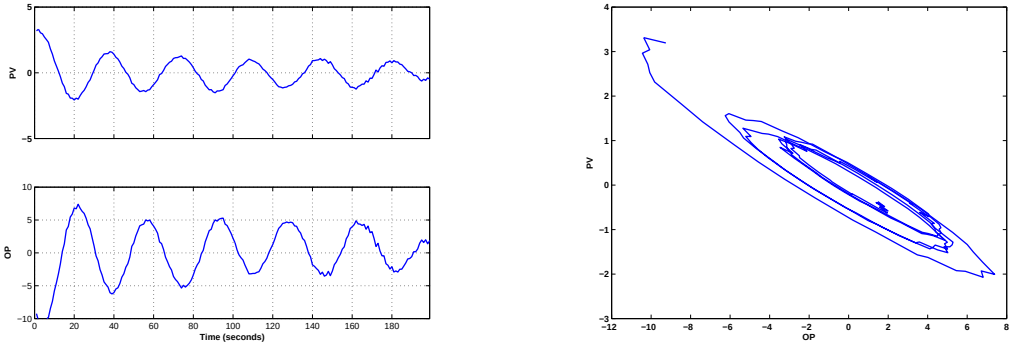
Control system	Case number	SI_{OP}
Flow control	Case 1 & Case 2	0.17
Flow control	Case 1 & Case 3	0.75

of oscillation, the calculated SI may not be able to clearly indicate whether there is a valve stiction or not. However, SI tells us the dominant factor which cause the oscillation. For example, in the flow control, if Case 1 and Case 2 exist simultaneously, or Case 1 and Case 3 exist simultaneously, the results are shown in Table 3.4. As we can see, for the case where both external disturbance and weak stiction exist, the SI indicates the dominant factor is external disturbance while for the case where both external disturbance and strong stiction exist, the SI indicates the dominant factor is valve stiction.

3.6 Industrial Examples

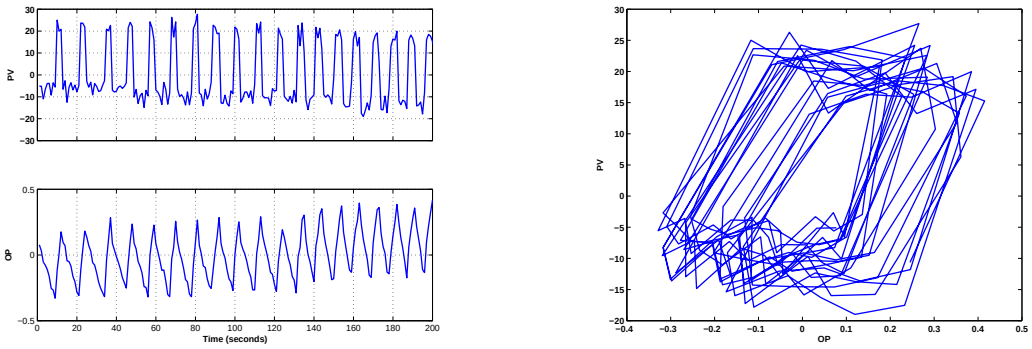
Three cases from chemical processes are investigated in this section: case 1 is a temperature control loop which is over aggressively tuned; case 2 is a flow control loop and it is known that this loop has valve stiction problem; case 3 is a level control loop which also has valve stiction problem. Figures 3.17, 3.18 and 3.19 show normalized operation data and Table 3.5 summarizes detection results by stiction indices[3]. In all three cases, the proposed method successfully detects valve stiction.

[3]Stiction index for case 3 is based on process output while indices for case 1 and 2 are based on controller outputs.



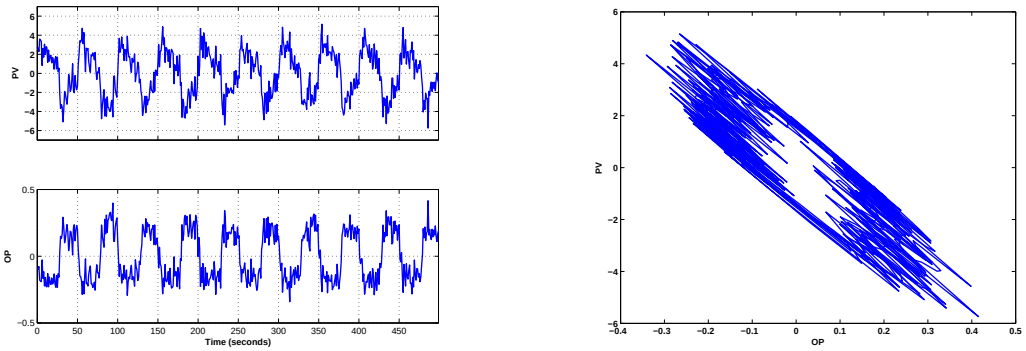
(a) (b)

Figure 3.17: Industrial example, case 1 – temperature control with aggressive tuning



(a) (b)

Figure 3.18: Industrial example, case 2 – flow control with valve stiction



(a) (b)

Figure 3.19: Industrial example, case 3 – level control with valve stiction

Table 3.5: Industrial examples

Control system	Case number	SI
Temperature control	Case 1	0.20
Flow control	Case 2	0.69
Level control	Case 3	0.94

3.7 Conclusions

In this work, an existing valve stiction model is reviewed and its shortfall is discussed. A structurally simple and logically straightforward valve stiction model is presented. One valve stiction detection method proposed by Horch is examined and its inconsistence is theoretically analyzed and demonstrated by a simple example. A simple and effective valve stiction detection method is proposed. The method is to fit two different functions, triangle and sinusoid, to the measured controller output for self-regulating processes (or process output for integrating processes). A better fit to the triangle indicates valve stiction, a better fit to the sinusoid, non-stiction. Stiction index as a new

metric is defined and used as a criterion to estimate the probability of valve stiction. The method is evaluated on simulated examples and industrial data sets where the actual oscillation causes are known. The proposed method and metric are shown to successfully detect valve stiction in both self-regulating and integrating processes and distinguish valve stiction from external oscillatory disturbance and bad tuning.

Chapter 4

A New Fault Diagnosis Method Using Fault Directions in Fisher Discriminant Analysis

Multivariate statistical methods such as principal component analysis (PCA) and partial least squares (PLS) have been widely applied to the statistical process monitoring (SPM) of chemical processes and their effectiveness for fault detection is well recognized. These methods make use of normal process data to define a tight normal operation region for monitoring. In practice, however, historical process data are often corrupted with faulty data. In this work, a new process monitoring method is proposed which is composed of three parts: (i) a pre-analysis step that first roughly identifies various clusters in a historical data set and then precisely isolates normal and abnormal data clusters by the k-means clustering method; (ii) a fault visualization step that visualizes high-dimensional data in 2-D space by performing global Fisher discriminant analysis (FDA), and (iii) a new fault diagnosis method based on fault directions in pair-wise FDA. A simulation example is used to demonstrate the performance of the proposed fault diagnosis method. An industrial film process is used to illustrate a realistic scenario for data pre-analysis, fault visualization and fault diagnosis. In both examples, the contribution plots method based on fault directions in pair-wise FDA shows superior capability

for fault diagnosis to the contribution plots method based on PCA.

4.1 Introduction

As chemical processes become more complex, the monitoring of chemical processes is gaining importance in order to assess process performance and improve process efficiency and product quality. Early detection of faults can help avoid major breakdowns and incidents. In general, four tasks are involved in the process monitoring: fault detection, which gives an indication that something is going wrong in the process; fault identification or diagnosis, which determines the root cause of the fault; fault estimation, which assesses the size of the fault; and fault reconstruction, which estimates the fault-free values [80]. Traditional fault detection and isolation (FDI) methods have been based on a mathematical model of the system. These approaches make use of the state estimation, parameter identification techniques, and parity relations to generate residuals [7, 31, 59]. However, it is often difficult and time-consuming to develop accurate mathematical models that characterize all the physical and chemical phenomena occurring in industrial processes. Knowledge-based approaches such as expert systems may be considered as alternative or complementary approaches to the analytical model-based approaches where analytical models are not available [27]. However, considerable effort is also required to build these knowledge-based systems [104].

In order to address the difficulties that lie in the model-based or knowledge-based methods, model-free statistical process monitoring (SPM) methods have

been developed. SPM methods only require a good historical data set of normal operations which is available for computer controlled industrial processes. Due to the data-based nature of the SPM methods, it is relatively easy to apply to rather large and complex processes comparing to model-based or knowledge-based approaches. The traditional univariate statistical process control (SPC) charts, such as the Shewhart chart, CUSUM plot and EWMA chart, are well established statistical procedures for monitoring stable processes. While univariate statistical techniques are easy to implement, they often lead to significant number of false alarms on multivariate chemical processes where the sensor measurements are highly correlated due to physical and chemical principles governing the process operation, such as mass and energy balances [22]. A simple yet illustrative example which shows the misleading nature of the univariate charts is given by Kourti and MacGregor [67] where the true situation is only revealed in a multivariate plot. Multivariate statistical process control charts based on multivariate statistical methods, such as principal component analysis (PCA) and partial least squares (PLS), have been developed to overcome the shortcomings of univariate SPC.

In PCA or PLS based process monitoring, two indices have been widely used for fault detection: the Hotelling's T^2 statistic which gives a measure of the variation with the PCA model, and the squared prediction error (SPE) of the residuals which indicates how much each sample deviates from the model. Other less commonly used indices, such as Hawkins' T_H^2 [36], Mahalanobis distance and combined indices [82, 105, 106] have been proposed and their pros

and cons are discussed in [80, 93].

After a fault has been detected, fault diagnosis becomes important because it is desirable to find the root cause of the fault. Currently, the well known fault diagnosis approaches based on PCA and PLS models are the contribution plots and reconstruction based methods [80]. Contribution plots are very easy to generate with no prior process knowledge. Contribution plots show the contribution of each process variable to the observed statistic, i.e. SPE or T^2. It is assumed that the process variable with high contribution is likely the root cause of the fault. However, the contribution plots may not explicitly identify the cause of an abnormal condition [66], and sometimes may lead to incorrect conclusions. One reason is that the contribution from one variable is propagated to other variables in calculating the projection. This 'smearing' effect can reduce the significance between contributing and non-contributing variables [80]. Due to limited redundancy or correlation among the process variables, it is possible that some faults may not be identifiable [80].

Furthermore, the PCA approach assumes that normal data have already been isolated from historical operational data. The reality is that historical data often contain both normal and abnormal data, but little work has been done to isolate normal data from abnormal data. In this work, we start with the assumption that the historical data may contain both normal and multiple classes of abnormal data. The first step of the approach is to visualize the number of classes in the data using PCA score plots, the SPE and T^2 charts. The historical data are then classified into different classes us-

ing k-means clustering. In the next step, we apply global Fisher discriminant analysis (FDA) to normal data and all classes of fault data to obtain a clear class visualization of high-dimensional data. In the last step, pair-wise FDA is applied to normal data and each class of fault data to find fault directions that optimally separate fault data from normal data. The weights in fault directions are used to generate contribution plots for fault diagnosis. The new approach is applied to the fault diagnosis of a simulation example, the quadruple tank process, and an industrial polyester film process. The results show that the pair-wise FDA provides an optimal set of fault directions in terms of distinguishing fault data from normal data and is shown to be superior for fault diagnosis compared to PCA based contribution plots. Furthermore, in the industrial example, the visualization of lower-dimensional representation in FDA Fisher space gives a clearer view in terms of maximizing the separation amongst multiple classes than that in the PCA score space.

It should be noted that FDA is a widely used technique in pattern classification [84], but its use for analyzing chemical process data has not been explored until recently [12, 13]. The basic idea of FDA is to find the Fisher optimal discriminant vector such that the Fisher criterion function is maximized. While PCA seeks directions that are efficient for representation, FDA seeks directions that are efficient for discrimination. Therefore, FDA has advantages for fault visualization and diagnosis from a theoretical point of view [12]. In this work, we develop a novel fault diagnosis approach based on fault directions in pair-wise FDA.

The organization of the chapter is as follows. Section 5.2 gives preliminaries which provide some background knowledge on PCA, FDA and k-means clustering. Section 5.3 introduces the new fault diagnosis method that includes data pre-analysis, fault visualization and fault diagnosis using fault directions defined by pair-wise FDA. A simulation example is given in Section 5.4 to demonstrates the advantage of the pair-wise FDA for fault diagnosis. Section 5.5 presents an application of pre-analysis, fault visualization and fault diagnosis to an industrial example. Section 5.6 gives conclusions to the chapter.

4.2 Preliminary

In this section, we briefly review some relevant methods for fault diagnosis and classification. The fault detection and diagnosis method based on PCA will be introduced first, then we will review FDA which is the basis of the proposed fault diagnosis method. Finally, we will introduce the k-means clustering method, which is used in the data pre-analysis.

4.2.1 PCA-based process monitoring

Principal component analysis in many ways forms the basis of multivariate data analysis [102]. Let $X^0 \in \Re^{n\times m}$ denote the raw data matrix with n samples (rows) and m variables (columns). X^0 is first scaled to a matrix X with zero mean for covariance-based PCA and, with unit variance for correlation-based PCA. By either the NIPALS [102] or a singular value decomposition (SVD) algorithm, the scaled matrix X is decomposed as follows:

$$X = TP^T + \tilde{X} = TP^T + \tilde{T}\tilde{P}^T = \left[T\ \tilde{T}\right]\left[P\ \tilde{P}\right]^T \tag{4.1}$$

where $T \in \Re^{n\times l}$ and $P \in \Re^{m\times l}$ are the score matrix and the loading matrix, respectively. The PCA projection reduces the original set of m variables to l principal components. The decomposition is made such that $[T\ \tilde{T}]$ is orthogonal and $[P\ \tilde{P}]$ is orthonormal. The columns of P are actually eigenvectors of the covariance or correlation matrix of the variables associated with the l largest eigenvalues, and the columns of $\tilde{P}$ are the remaining eigenvectors. For fault detection in a new sample vector x, the squared prediction error (SPE)

and the Hotelling's T^2 are often used. The SPE statistic indicates how well each sample conforms to the model, measured by the projection of the sample vector on the residual space:

$$SPE = \| \tilde{x} \|^2 = \| \left(I - PP^T \right) x \|^2 \tag{4.2}$$

The process is considered normal if

$$SPE \leq \delta_\alpha^2 \tag{4.3}$$

where δ_α^2 denotes the upper control limit for SPE with a significance level α. An expression for δ_α^2 has been developed by Jackson and Mudholkar [60] assuming that x follows a normal distribution.

The Hotelling's T^2 is a measure of the variation in principal component space:

$$T^2 = x^T P \Lambda^{-1} P^T x \tag{4.4}$$

The T^2 statistic forms an ellipse, which represents the joint limits of variations that can be explained by a set of common causes. For a given significance level α, the process is considered normal if

$$T^2 \leq T_\alpha^2 \tag{4.5}$$

where the upper control limit T_α^2 can be calculated or approximated in several ways [80]. If both process data X and quality data Y are available and one wishes to extract variations in X that contribute to Y, PLS should be used instead of PCA. PLS attempts to extract the latent variables that not only

explain the variations in the process data X, but also the variations in X which are more predictive of the quality data Y. Since only the process data will be used in this work, PLS will not be discussed. Interested readers should refer to [67, 69, 101].

The SPE and Hotelling's T^2 are adequate to detect when the process is out-of-control, but they cannot indicate which variables are responsible for the malfunction. The contribution plots are well known tools for fault diagnosis [66, 68, 71, 72, 75], which break down the SPE or T^2 into each element corresponding to the contribution from each variable. The contribution for SPE is simply breaking down the SPE into each element:

$$SPE = \sum_{i=1}^{m} \tilde{x}_i^2 \tag{4.6}$$

where $\tilde{x}_i$ is the contribution to SPE from the i^{th} variable. If a sample has an abnormal SPE, the variables with the largest contributions are investigated.

The contribution plot on PCA scores indicates how significant is the effect of each variable on the T^2. The variables with the largest contributions are considered major contributors to the fault. The T^2 contribution can be defined in several ways [75, 76, 79, 99]. Upper control limits for contribution plots are discussed in [16, 79, 99].

4.2.2 Fisher discriminant analysis

Fisher discriminant analysis is a widely used technique in pattern classification. The basic idea of FDA is to find the Fisher optimal discriminant

vector such that the Fisher criterion function is maximized. The higher-dimensional feature space then can be projected onto the obtained optimal discriminant vectors for constructing a lower-dimensional feature space. Let $X \in \Re^{n \times m}$ be a set of m-dimensional samples $x \in \Re^m$ and the matrix X_i is the subset containing n_i rows of X corresponding to the samples from class i. If $\bar{x}_i$ is the m-dimensional sample mean for class i given by

$$\bar{x}_i = \frac{1}{n_i} \sum_{x \in X_i} x \tag{4.7}$$

then the within-class scatter matrix is given by

$$S_w = \sum_{i=1}^{c} P(\omega_i) S_i \tag{4.8}$$

where

$$S_i = \frac{1}{n_i} \sum_{x \in X_i} (x - \bar{x}_i)(x - \bar{x}_i)^T \tag{4.9}$$

is the within-class scatter matrix for class i and $P(\omega_i)$ is the a priori probability of class i, generally, $P(\omega_i) = 1/c$.

Let $\bar{x}$ be the mean vector of all samples in X, the between-class scatter matrix is defined by

$$S_b = \sum_{i=1}^{c} P(\omega_i)(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T \tag{4.10}$$

The optimal discriminant direction is found by maximizing the Fisher criterion:

$$J(\varphi) = \frac{\varphi^T S_b \varphi}{\varphi^T S_w \varphi} \tag{4.11}$$

where the maximizer φ is the Fisher optimal discriminant direction which maximizes the ratio of the between-class scatter to the within-class scatter. It is easy to show that a vector φ that maximize $J(\cdot)$ must satisfy

$$S_b\varphi = \lambda S_w\varphi \tag{4.12}$$

for some constant λ, which is a generalized eigenvalue problem. If S_w is non-singular, we can obtain a conventional eigenvalue problem by writing

$$S_w^{-1}S_b\varphi = \lambda\varphi \tag{4.13}$$

4.2.3 k-means clustering

Industrial data usually contain both normal and abnormal data in high dimensional space, making it difficult to segregate manually. In this work, k-means clustering is used to isolate different classes of data. k-means clustering can best be described as a partitioning method which partitions the samples in the data set into mutually exclusive clusters. Unlike the hierarchical clustering methods, k-means clustering does not create a tree structure to describe the groupings in the data set, but rather creates a single level of clusters. Compared to hierarchical clustering methods, k-means is more effective for clustering large amounts of data. The number of clusters, k, needs to be determined at the onset. The idea behind k-means clustering is to divide the samples into k clusters such that some metric relative to the centroids of the clusters is minimized. Various metrics to the centroids that can be minimized include:

- maximum distance to its centroid for any sample
- sum of the average distance to the centroids over all clusters
- sum of the variance over all clusters
- total distance between all samples and their centroids

The metric to minimize and the choice of a distance measure will determine the shape of the optimum clusters.

Suppose we are given $X \in \Re^{m \times n}$, a set of m samples in n-dimensional space $\Re^n$, and an integer k, and the problem is to determine a set of k centroids $\mu_1, \mu_2, \cdots, \mu_k$ in $\Re^n$, so as to minimize the sum-of-squares criterion

$$J = \sum_{c=1}^{k} \sum_{j=1}^{n} \|x_j - \mu_c\|^2 \tag{4.14}$$

A general algorithm is:

1. Randomly pick k samples in the data set as the initial cluster centroids $\mu_1^{(0)}, \mu_2^{(0)}, \cdots, \mu_k^{(0)}$. Set iteration $i = 0$.
2. Assign each sample x_j to the cluster with the nearest centroid $\mu_c^{(i)}$.
3. When all samples have been assigned, recalculate the positions of the k centroids
$$\mu_c^{(i+1)} = E\{x_j\}_{x_j \in \mu_c^{(i)}} \tag{4.15}$$
4. Repeat Steps 2 and 3 until the centroids no longer move.

The above k-means algorithm uses an iterative procedure which converges to one of the local minima. The computational complexity is $O(mnkT)$ where T is the number of iterations. In practice, the number of iterations is generally much less than the number of samples [84]. It is known that k-means are sensitive to initial starting conditions. Despite this limitation, the algorithm is used fairly frequently as a result of its ease of implementation. One way to find good optima is to do many runs of k-means, each from a different random starting points, and find the best minimum in terms of (4.14).

4.3 Fault Diagnosis Using Fault Directions in FDA

In this section, we present our proposed approach in three steps: data pre-analysis, fault visualization, and fault diagnosis. We start with the assumption that the historical data contain unclassified normal and multiple classes of abnormal data. By incorporating process knowledge, the first step of the approach is to visualize the number of classes in the data using PCA score plots, SPE chart and T^2 chart, the historical data are then classified into different classes using k-means clustering. In the next step, global FDA is applied to obtain a clear fault visualization in 2-D or 3-D Fisher space. Finally, pair-wise FDA is applied to normal data and each class of fault data to find fault direction that optimally separates each fault data from normal data. The weights in fault directions are used to generate contribution plots for fault diagnosis. The entire process, including pre-analysis of historical data, fault visualization and fault diagnosis, is summarized in Figure 4.1 and each step is discussed in the following subsections.

4.3.1 Data pre-analysis

In this step, a PCA model is built for the whole data set which contains both normal and abnormal data. Clusters are visualized in PCA 2-D or 3-D score space. The total number of clusters k can usually be revealed in the PCA score plot, although the clusters are not maximally separated. Then a finer analysis that incorporates process knowledge is conducted to determine the total number of clusters k in the data. The objective is to identify enough

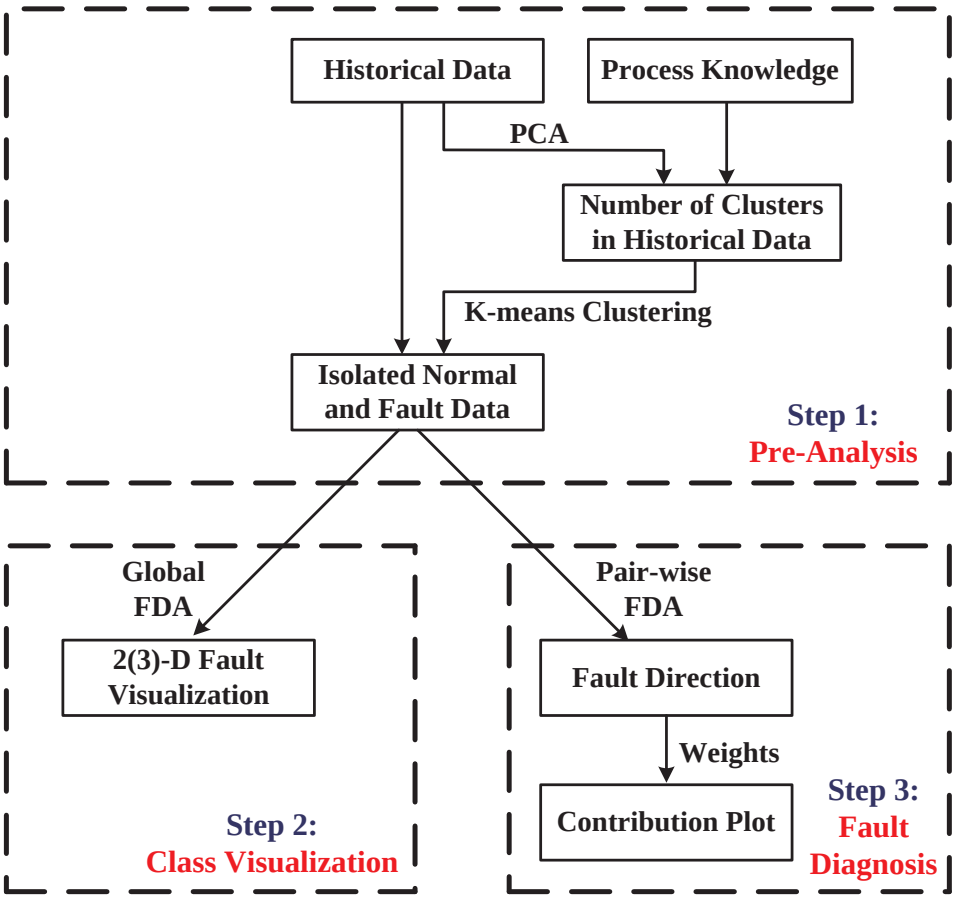


Figure 4.1: Overall flow chart of the proposed pre-analysis, fault visualization and fault diagnosis method

normal samples to build the normal PCA model. Ambiguous data points during the transition between clusters can be ignored. After k is identified, X is partitioned into k disjoint classes: $X_0, X_1, \cdots, X_{k-1}$ by applying k-means clustering. This step usually involves several iterations by incorporating process knowledge.

4.3.2 Fault visualization

Data visualization is an active research field in computer science and is a desirable feature for process engineers to perform fault diagnosis. Although the data are high dimensional, it is possible to project the fault classes to low dimensional space using dimension reduction techniques such as PCA and FDA.

In this step, we apply global FDA to all classes identified in the first step. The within-class and between-class scatter matrices are calculated by Equations (4.8) and (4.10). Similar to the score plot based on PCA, we project high-dimensional data on to φ_1 and φ_2, corresponding to the first two largest eigenvalues λ_1 and λ_2, to obtain 2-D visualization of normal and fault data in FDA Fisher space. Because of the discrimination nature of FDA, we will have a better fault visualization than that in the PCA score space. This will be demonstrated in an industrial example later in this chapter.

4.3.3 Fault diagnosis

After the normal and fault data are properly classified, the next step is to characterize faults by pair-wisely applying FDA to normal data, denoted as X_0, and each class of fault data X_i $(i = 1, \cdots, k-1)$. The scatter matrix for normal data X_0 is

$$S_0 = \frac{1}{n_0} \sum_{x \in X_0} (x - \bar{x}_0)(x - \bar{x}_0)^T \tag{4.16}$$

The scatter matrix for fault data X_i $(i = 1, \cdots, k-1)$ is

$$S_i = \frac{1}{n_i} \sum_{x \in X_i} (x - \bar{x}_i)(x - \bar{x}_i)^T \tag{4.17}$$

Therefore the within-class scatter matrix is

$$S_w = S_0 + S_i \tag{4.18}$$

Let $\bar{x}$ be the mean vector of samples in X_0 and X_i, the between-class scatter matrix is given by

$$S_b = \sum_{j=0,i} (\bar{x}_j - \bar{x})(\bar{x}_j - \bar{x})^T \tag{4.19}$$

Since we include only two classes in pair-wise FDA analysis, by substituting Equations (4.18) and (4.19) into Equation (4.12) and solving the generalized eigenvalue problem we will obtain only one significant eigenvalue λ_i and one Fisher direction, i.e., the eigenvector φ_i, corresponding to this single significant λ_i. This Fisher direction is the optimal direction which discriminates fault data X_i from normal data X_0 according to the Fisher criterion. This direction best characterizes the effect of the fault relative to the normal data. Therefore, we

define this Fisher direction φ_i as the fault direction for X_i. The weights in φ_i are used to generate the contribution plot for fault X_i. For a fault direction

$$\varphi_i = [\phi_1, \phi_2, \cdots, \phi_j, \cdots, \phi_m]^T \tag{4.20}$$

the j^{th} element ϕ_j is the contribution from the j^{th} variable. Note that ϕ_j represents an average contribution over n_i samples in X_i because the fault direction is calculated based on all samples in X_i. The new fault diagnosis method is illustrated in Figure 4.2 where we start with isolated normal and fault data obtained from data pre-analysis. The fault direction is calculated by performing pair-wise FDA on normal and each class of fault data, then we examine the contribution plot based on the fault direction to determine the root cause of the fault. This process is repeated until all faults are analyzed.

A simple illustrative example is used here to demonstrate the procedure of finding fault directions and creating contribution plots using fault directions in pair-wise FDA. The data are generated in the following way using Matlab:

1. Normal samples (x_0, y_0):

$$\begin{aligned} x_0 &= \text{randn}(200, 1); \\ y_0 &= 0.7 * \text{randn}(200, 1); \end{aligned}$$

2. Fault samples (x_1, y_1) with bias in x direction:

$$\begin{aligned} x_1 &= x_0 + 6 + \text{noise}; \\ y_1 &= y_0 + \text{noise}; \end{aligned}$$

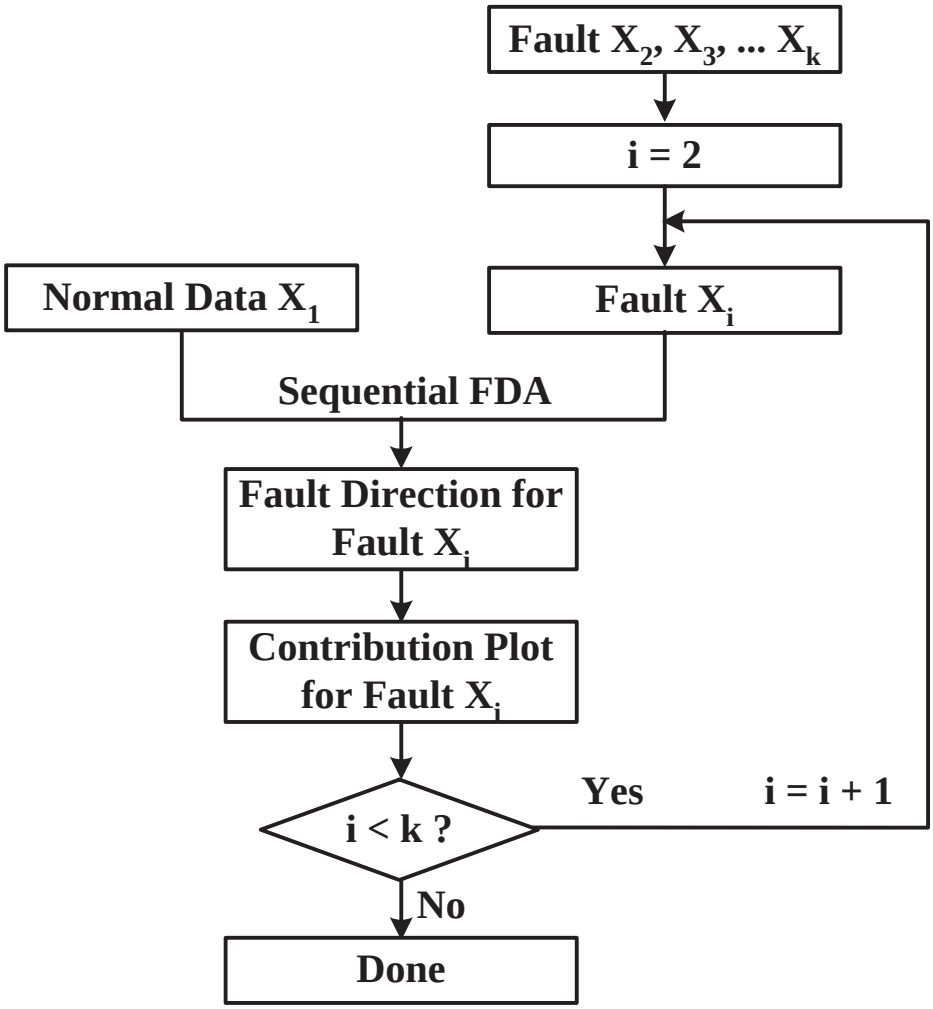


Figure 4.2: Pair-wise FDA flow chart

Figure 4.3(a) shows the scatter plot of the samples where stars are normal samples and x-mark's are fault samples with a mean shift in x direction. By performing FDA on normal and fault data, we find fault direction $\varphi_1 = [-0.00891 \;\; 0.99996]^T$ which is shown in Figure 4.3(a) as black dash line. For comparison, PCA direction is shown as gray dash line and PCA residual direction is shown as gray solid line in the same plot. The contribution plot is shown as the bar chart of φ_1 in Figure 4.3(b) where the gray bar is the contribution from x while the black bar is the contribution from y. Also included in Figure 4.3(b) are averaged contributions to SPE and T^2 based on PCA. We observe that contribution plot based on FDA and PCA T^2 correctly indicate that a fault in x direction is the root cause of the fault while PCA SPE based contribution plot leads to the opposite conclusion. Now we look at another situation where we have a bias fault in y direction instead of in x direction:

3. Fault samples (x_2, y_2) with bias in y direction are generated as follows:

$$\begin{aligned} x_2 &= x_0 + \text{noise}; \\ y_2 &= y_0 + 6 + \text{noise}; \end{aligned}$$

The scatter plot is shown in Figure 4.3(c). It can be seen in Figure 4.3(d) that the contribution plot based on PCA T^2 gives incorrect conclusion while contribution plots based on PCA SPE gives correct fault direction which is similar to fault direction in FDA.

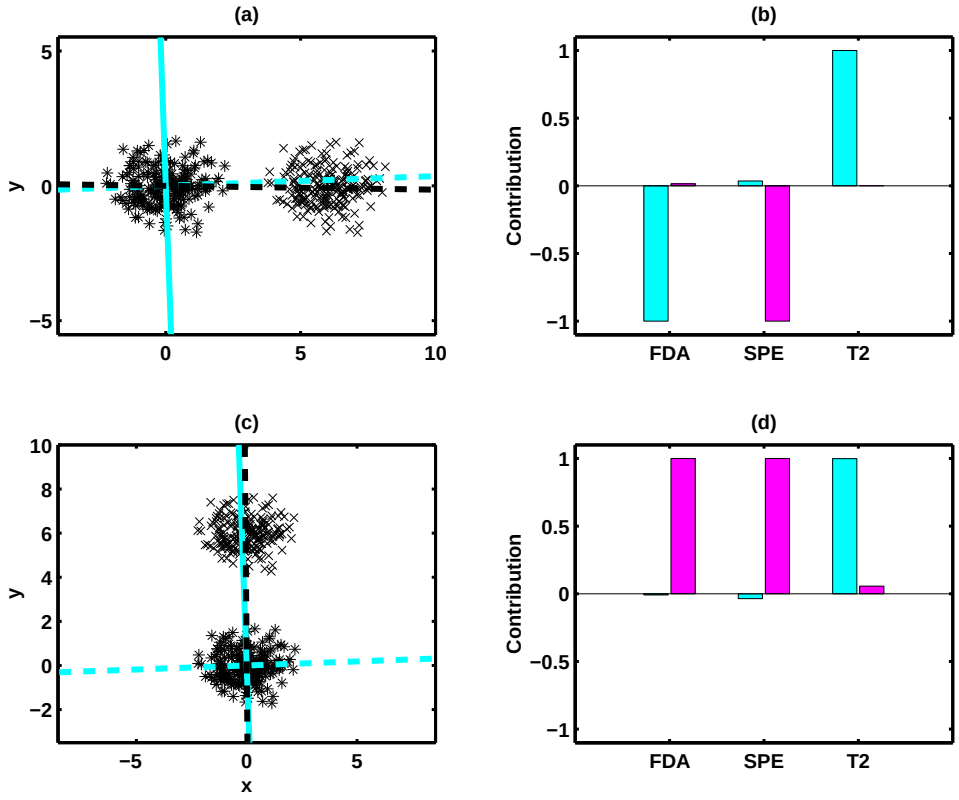


Figure 4.3: (a) Scatter plot – case 1; (b) Contribution plots – case 1; (c) Scatter plot – case 2; (d) Contribution plots – case 2

To summarize, contribution plots based on PCA T^2 and SPE do not give consistent conclusions for these two cases while contribution plots based on FDA fault directions give correct and consistent conclusions. Here we did not scale the variance of the data. If the normal data was scaled to unit variance, the fault direction angle based on PCA would be 45° and the contributions would be the same for two variables - no matter what kind of fault occurred to the fault data [80].

In the next two sections, the new approach is applied to the fault diagnosis of a simulation example, quadruple tank process, and an industrial polyester film process.

4.4 Simulation Example

In this section, the quadruple-tank process is used as a simulation example to demonstrate the advantage of the pair-wise FDA for fault diagnosis. The quadruple-tank process was originally developed by Johansson [61] as a novel multivariate laboratory process. This process consists of four interconnected water tanks, two pumps and associated valves. A schematic diagram of the process is shown in Figure 4.4. The inputs are the voltages supplied to the pumps, v_1 and v_2, and the outputs are the water levels $h_1 \sim h_4$. The flow to each tank is adjusted using the associated valves γ_1 and γ_2.

A nonlinear model is derived based on mass balances and Bernoulli's law:

$$\frac{dh_1}{dt} = -\frac{a_1}{A_1}\sqrt{2gh_1} + \frac{a_3}{A_1}\sqrt{2gh_3} + \frac{\gamma_1 k_1}{A_1}v_1 \tag{4.21}$$

$$\frac{dh_2}{dt} = -\frac{a_2}{A_2}\sqrt{2gh_2} + \frac{a_4}{A_2}\sqrt{2gh_4} + \frac{\gamma_2 k_2}{A_2}v_2 \tag{4.22}$$

$$\frac{dh_3}{dt} = -\frac{a_3}{A_3}\sqrt{2gh_3} + \frac{(1-\gamma_2)k_2}{A_3}v_2 \tag{4.23}$$

$$\frac{dh_4}{dt} = -\frac{a_4}{A_4}\sqrt{2gh_4} + \frac{(1-\gamma_1)k_1}{A_4}v_1 \tag{4.24}$$

For tank i, A_i is the cross-section of the tank, a_i the cross-section of the outlet hole, and h_i the water level. The voltage applied to pump i is v_i and the corresponding flow is $k_i v_i$. The parameters $\gamma_1, \gamma_2 \in (0, 1)$ are determined from how the valves are set prior to an experiment. The water flow rate to tank 1, i.e. f_1, is $\gamma_1 k_1 v_1$ and the flow rate to tank 4, i.e. f_4, is $(1-\gamma_1)k_1 v_1$ and similarly for tanks 2 and 3. The acceleration of gravity is denoted as g. The

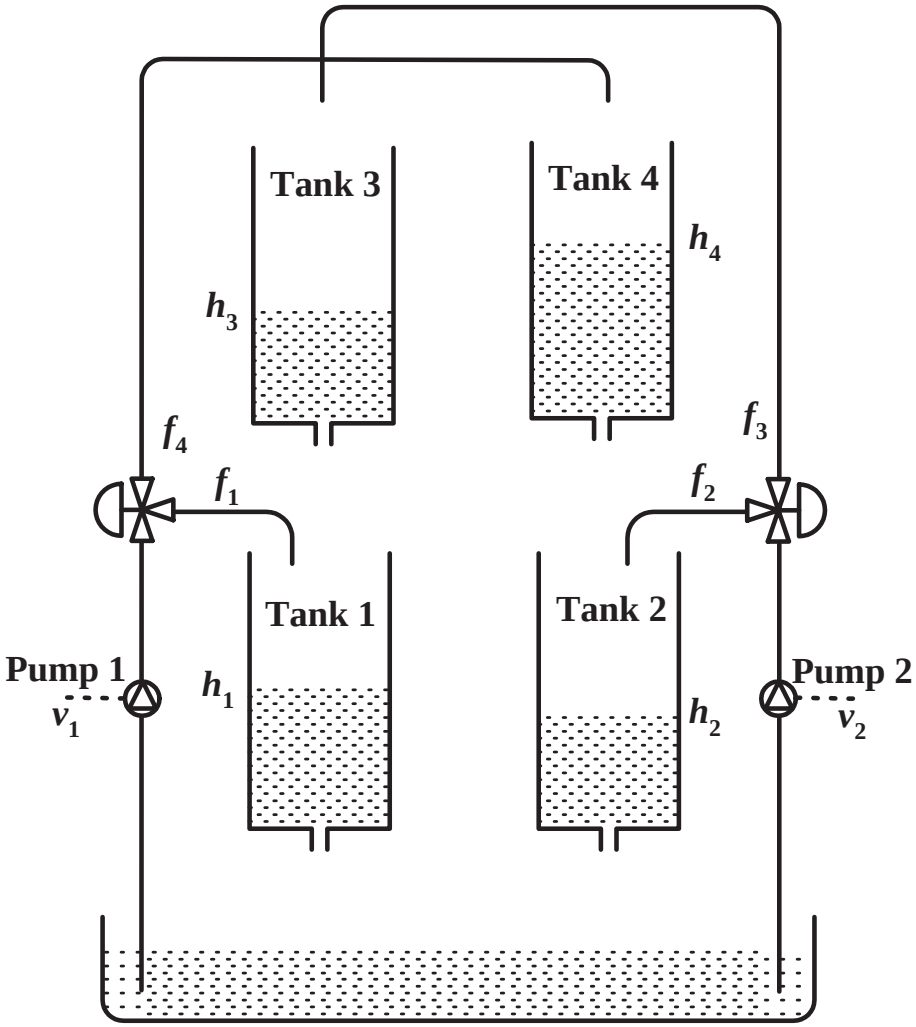


Figure 4.4: Schematic diagram of the quadruple-tank process

Table 4.1: Simulation parameters [61]

Parameter	Unit	Values
A_1, A_3	(cm^2)	28
A_2, A_4	(cm^2)	32
a_1, a_3	(cm^2)	0.071
a_2, a_4	(cm^2)	0.057
k_1, k_2	(cm^3/Vs)	3.33
g	(cm/s^2)	981

parameter values of this process are given in Table 4.1. The data is generated by Equations (4.21) $\sim$ (4.24), where γ_i and v_i are corrupted by independently Gaussian white noise with zero mean and standard deviation of 0.01 and 0.05 respectively, which are about $1 \sim 2\%$ of their upper limit or steady-state value. Measured h_i is corrupted by Gaussian distributed white noise with zero mean and standard deviation of 0.1 and measured γ_i and v_i contain the same level of noise as the input γ_i and v_i respectively.

Two cases, sensor fault and tank leakage, are studied in this work. In both cases, PCA T^2 chart and SPE chart are applied to detect the fault. Contribution plots based on both PCA and FDA fault directions are used to diagnose the fault and their performances are compared.

4.4.1 Case 1: sensor fault

We generate 100 normal data samples with sampling interval 10s, and then generate additional 100 samples with a bias fault $\Delta h_4 = 0.3$ in sensor h_4. Figure 4.5 shows the time series data of the process variables: water levels

$h_1 \sim h_4$ and flow rates to tanks $f_1 \sim f_4$. It is almost impossible for bare eyes to detect the sensor fault in h_4 from these plots due to the small magnitude of the bias. First, PCA is applied to analyze these data. A PCA model is built based on the first 100 observations and 4 PC's are kept in the model which capture about 76% of the total variance. The SPE and T^2 charts for normal and fault observations are given in Figure 4.6 with upper control limits. Both charts captured the subtle change in the process and clearly indicate that there is something went wrong after 1000s. Contribution plot is then created to diagnose the fault. The averaged contribution plot based on PCA SPE in Figure 4.7(a) does not explicitly indicate that h_4 is the root cause of the fault. Instead, both h_2 and h_4 are identified as the biggest contributors to this fault. FDA is then applied to diagnose the same fault. The FDA model is built based on the fault detection knowledge from PCA that there are two different classes of observations: the first 100 normal observations and the second 100 fault observations. The fault direction φ is found by solving Equation (4.12). The contribution plot based on FDA fault direction is shown in Figure 4.7(b) which clearly indicates that h_4 is the only root cause of the fault.

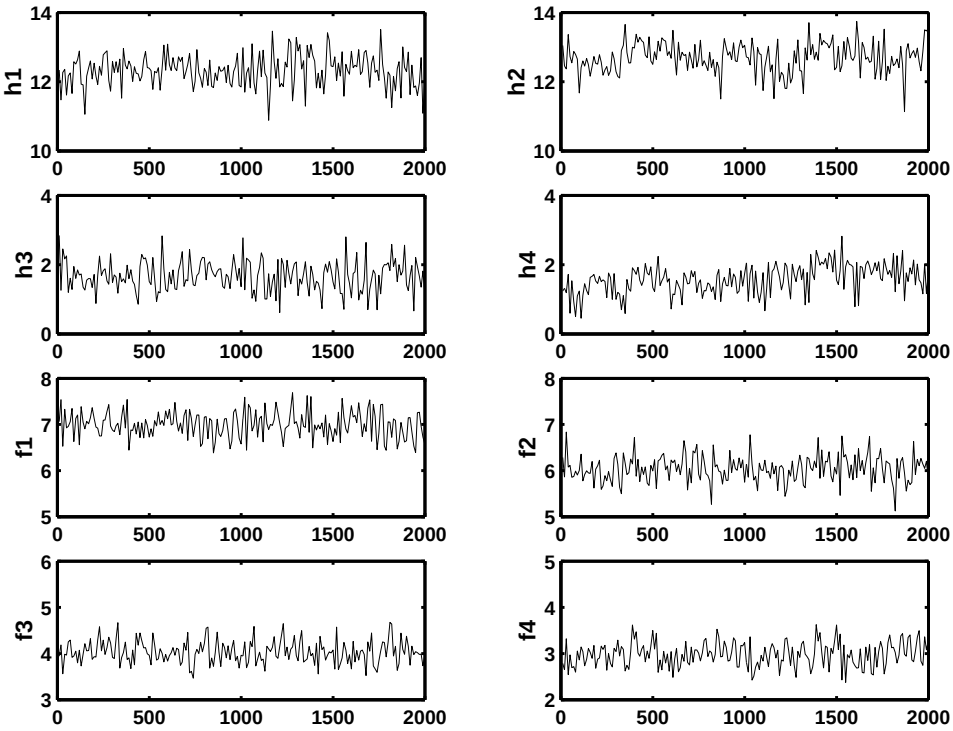


Figure 4.5: Process time series data with sensor fault in h_4

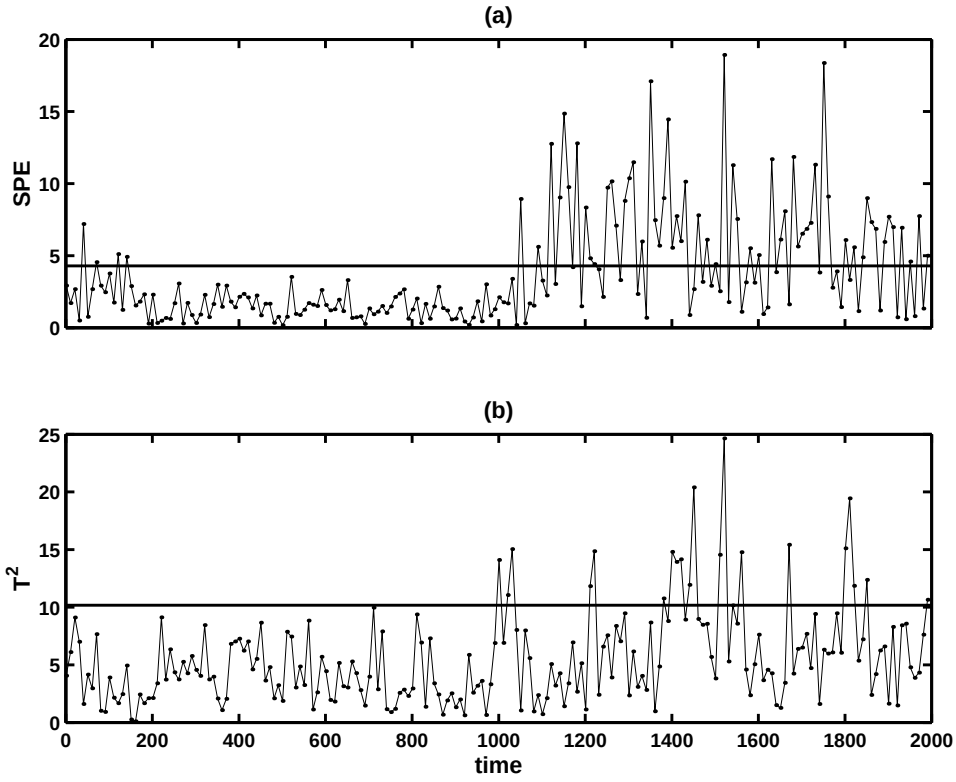


Figure 4.6: SPE and T^2 charts with 95% limit (the sensor fault in h_4 is introduced after 1000s)

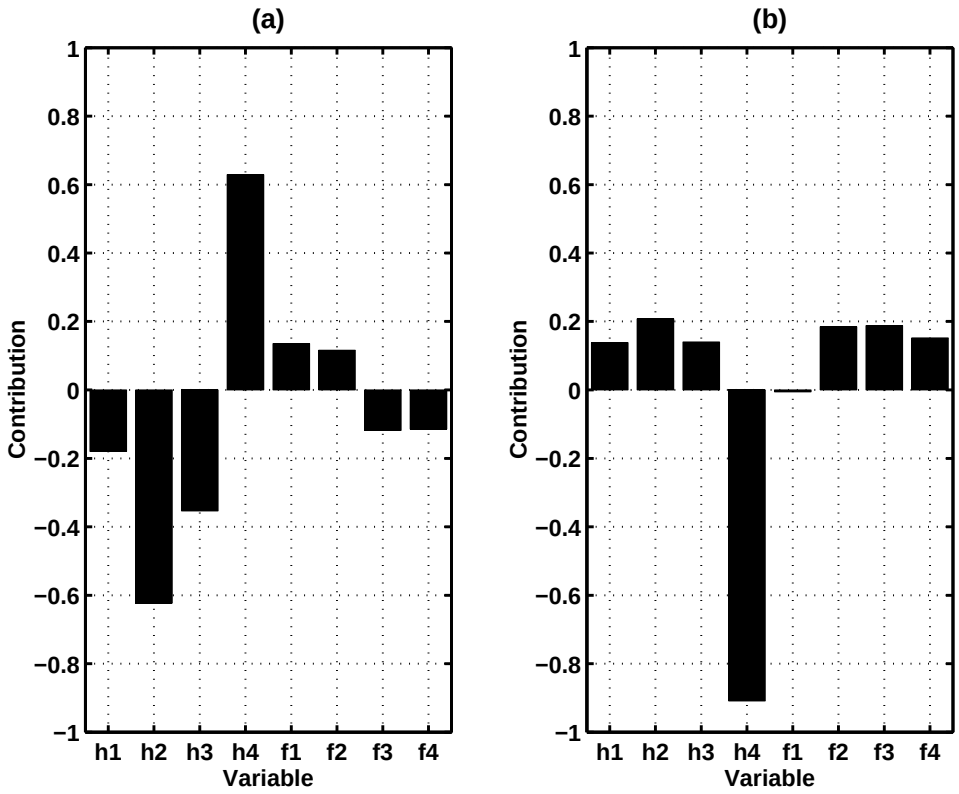


Figure 4.7: Contribution plots based on PCA model (a) and FDA fault direction (b) with sensor fault in h_4

4.4.2 Case 2: tank leakage

The data generation in case 2 is similar to case 1. A leakage in tank 1 is introduced after 1000s. We assume that there is a small hole at the bottom of tank 1 with the cross-section $a_{leak} = 0.005$ cm^2. The mass balance equation for tank 1 changes from Eqn. (4.21) to:

$$\frac{dh_1}{dt} = -\frac{a_1}{A_1}\sqrt{2gh_1} + \frac{a_3}{A_1}\sqrt{2gh_3} + \frac{\gamma_1 k_1}{A_1}v_1 - \frac{a_{leak}}{A_1}\sqrt{2gh_1} \tag{4.25}$$

where the last term corresponds to the leakage of tank 1. Mass balance equations for other tanks do not change. Figure 4.8 shows the time series data of the process variables consist of 100 normal data and 100 abnormal data. As in case 1, a PCA model is built based on normal data using 4 PC's. Both SPE chart and T^2 chart in Figure 4.9 detect the fault correctly. However, h_1 and h_3 are identified as the root cause of the fault as indicated in the contribution plot based on PCA SPE in Figure 4.10(a). Figure 4.10(b) shows the contribution plot based on the fault direction in FDA model, which is built based on the same procedure as in case 1. h_1 is explicitly identified as the root cause of the fault.

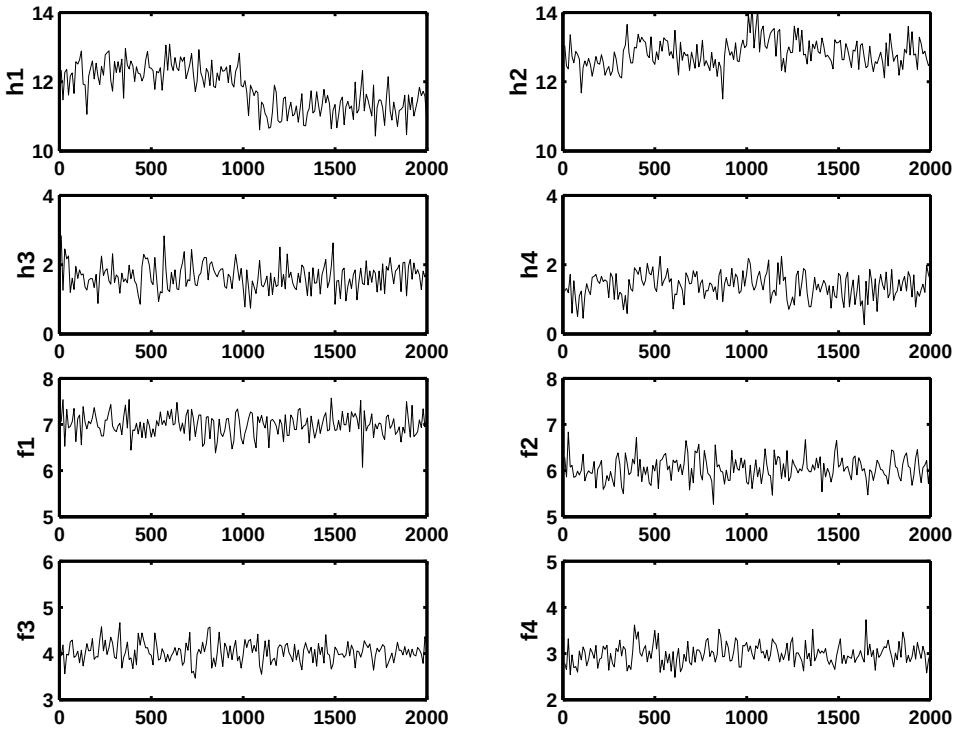


Figure 4.8: Process time series data with leakage in h_1

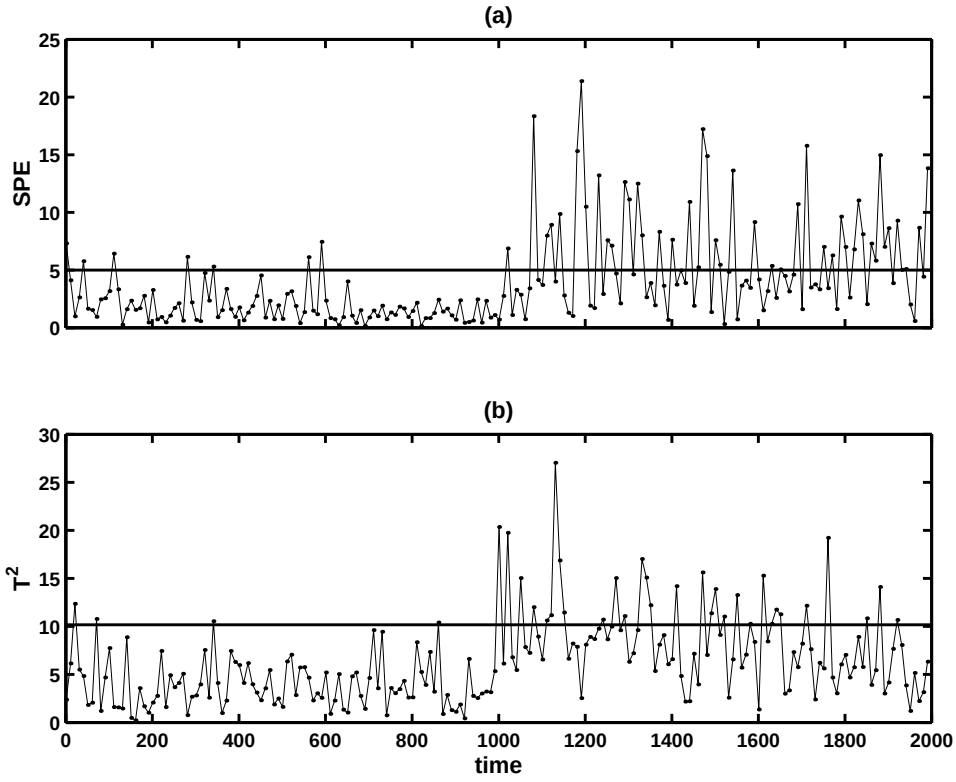


Figure 4.9: SPE and T^2 charts with 95% limit (the leakage in h_1 is introduced after 1000s)

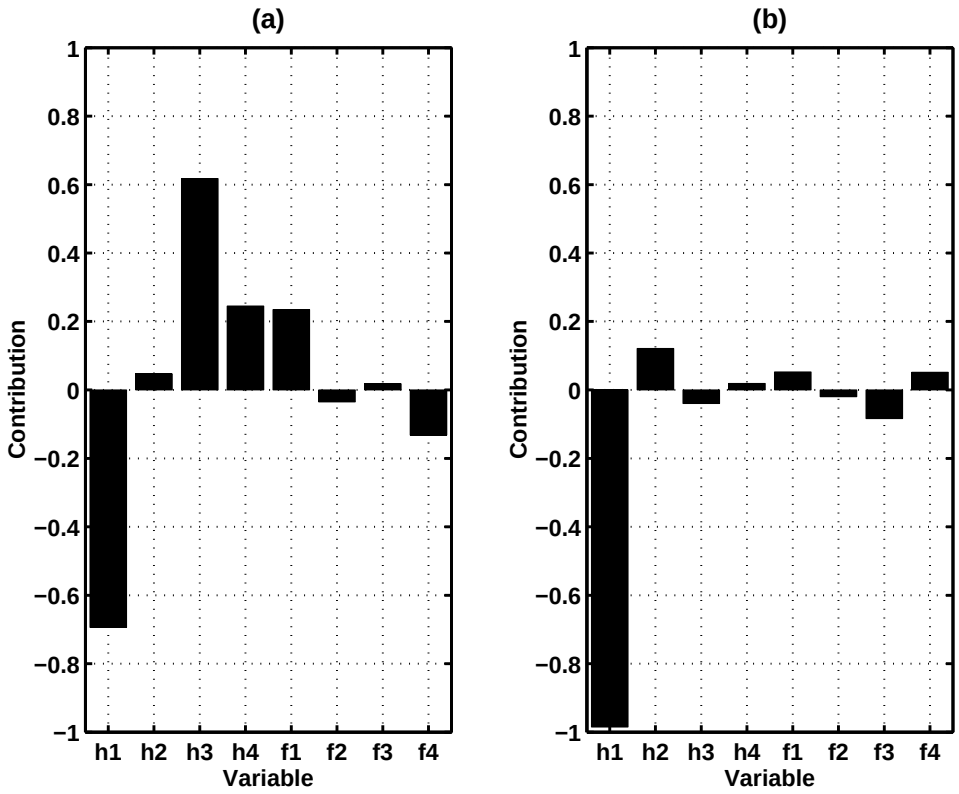


Figure 4.10: Contribution plots based on PCA model (a) and FDA fault direction (b) with leakage in h_1

4.5 Pre-analysis, Visualization and Diagnosis for an Industrial Film Process

In the simulation example, we only demonstrated the third step of the proposed approach - fault diagnosis using fault directions in FDA. Now we apply all three steps to an industrial polyester film manufacturing process. The process data contain a total of 2879 samples, which are mixture of normal and abnormal samples. Each sample consists of 103 measured process or monitoring variables belong to seven different operation zones, as shown in Table 4.2, to describe a unit or a specific physical or chemical operation [79]. This process

Table 4.2: Polyester film manufacturing process variables divided into blocks [79]

Block number	Process section	Variables in each block
1	Drying zone	1-9
2	Extrusion zone	10-29
3	Melt pipes zone 1	30-40
4	Melt pipes zone 2	41-52
5	Die zone	53-61
6	Casting zone	62-77
7	Tenter zone	78-103

is used to illustrate a realistic scenario for data pre-analysis, fault visualization, and fault diagnosis. In the first step, the k-means clustering method is used in conjunction with PCA based SPE chart and T^2 chart to classify the historical data into normal and abnormal operating regions. In the second step, global FDA is applied to visualize faults in 2-D Fisher space. In the third step, pair-wise FDA is applied to find fault directions that best isolate fault

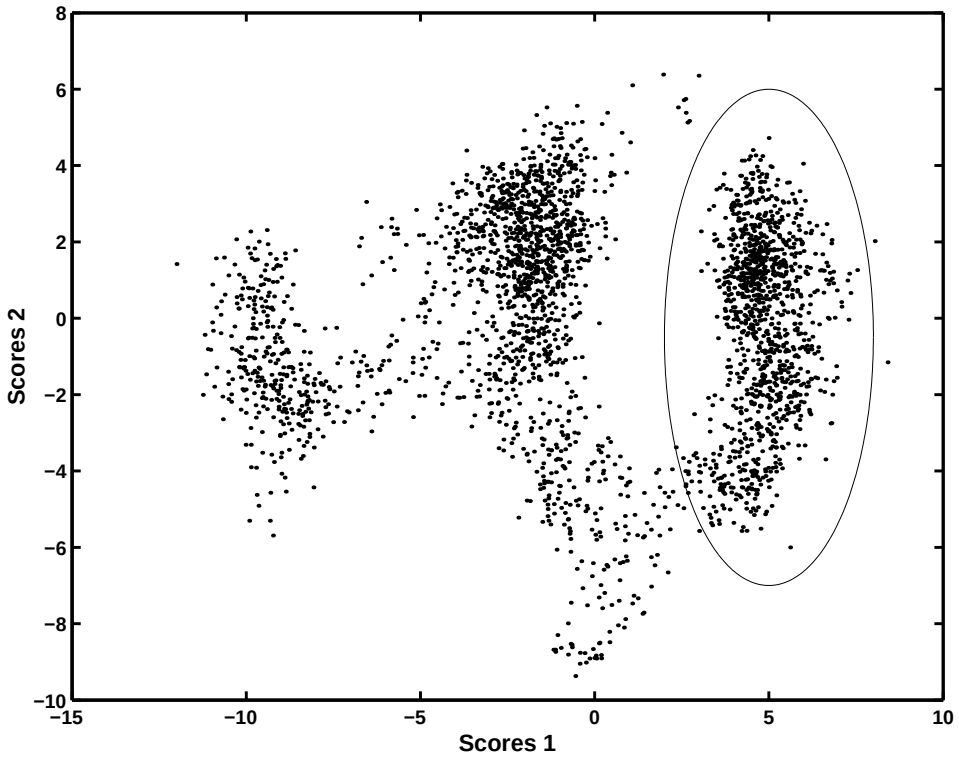


Figure 4.11: Clusters in the polyester film process data

data from normal data. These directions are interpreted as variable contributions for fault diagnosis and their performances are compared to PCA-based contribution plots.

4.5.1 Historical data pre-analysis

A preliminary PCA score plot based on all process data is shown in Figure 4.11 to visualize the clusters in the data set. We can see that there are several clusters. With the help from the plant engineers we know that the data cluster with an ellipse is normal while others are abnormal. From this plot it

is difficult to see whether there are two or three clusters outside the ellipse. So we rebuild PCA model based on the first 1000 normal process data only, then project the whole data set onto this PCA model. Figure 4.12(a) shows the SPE plot based on the PCA model and Figure 4.12(b) shows the Hotteling's T^2 plot. From these figures, we see clearly that there are four operation regions where A is the normal region, B, C, and D are fault regions. Figure 4.13 shows the PCA scores with approximately labelled regions based on SPE and T^2 charts in Figure 4.12. To determine the boundary of each region, k-means clustering is applied to obtain the exact category for each sample. Figure 4.14 shows the k-means clustering results in PCA scores space. The classes vs. samples are shown in Figure 4.15 where classes A, B, C and D correspond to the operation regions A, B, C and D.

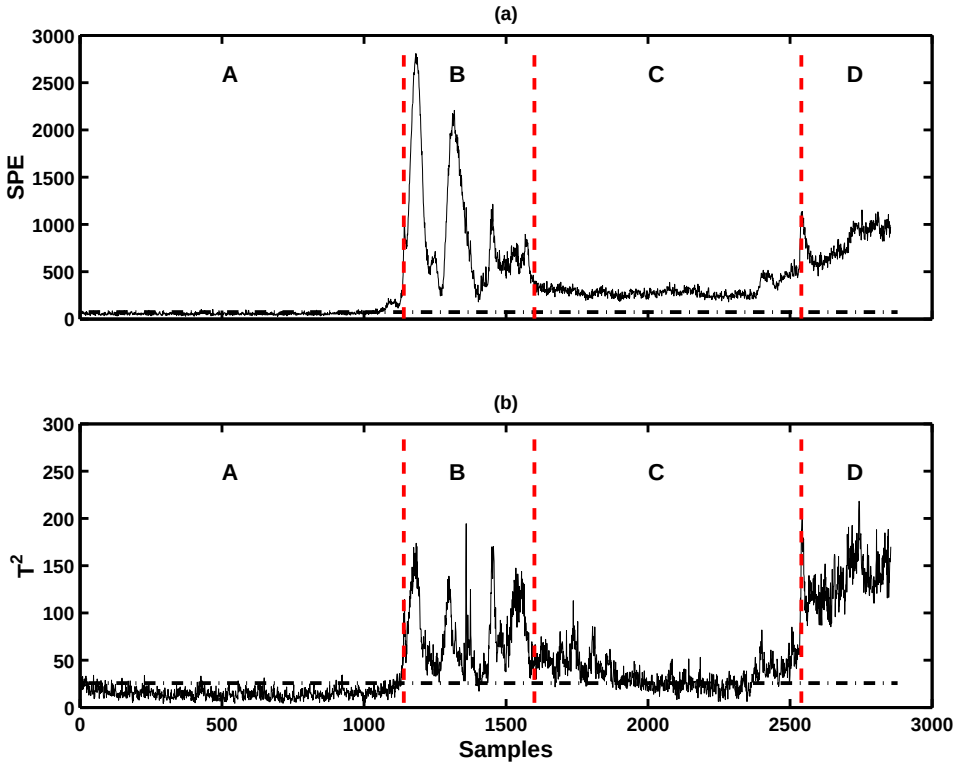


Figure 4.12: (a) SPE chart and (b) T^2 chart

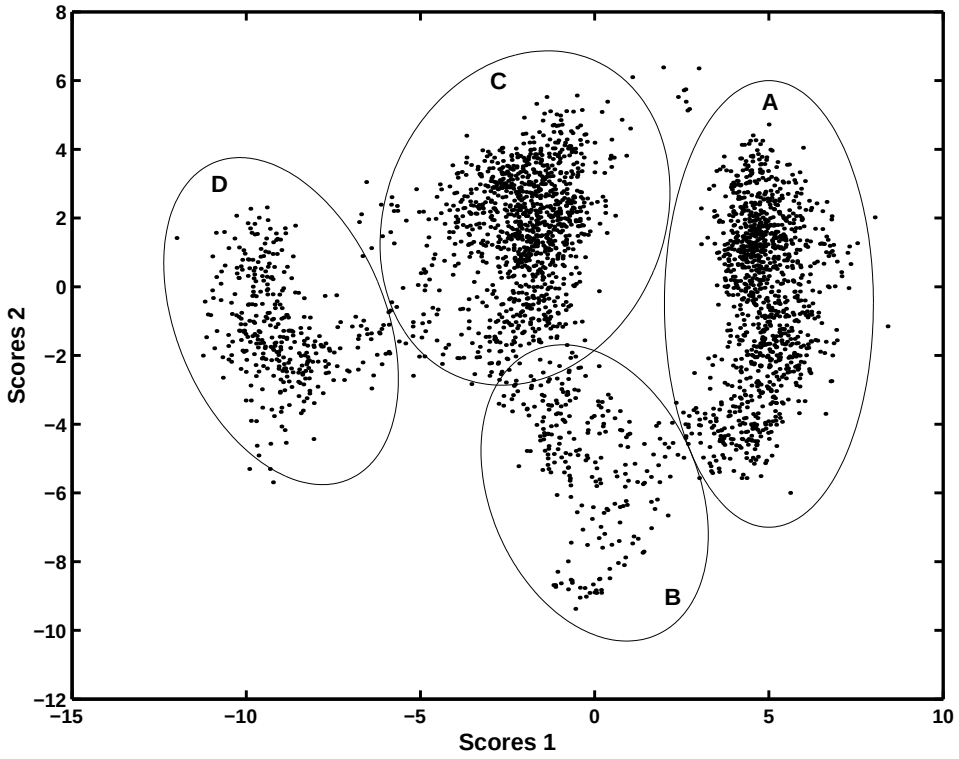


Figure 4.13: PCA approximately classified clusters in PCA score space

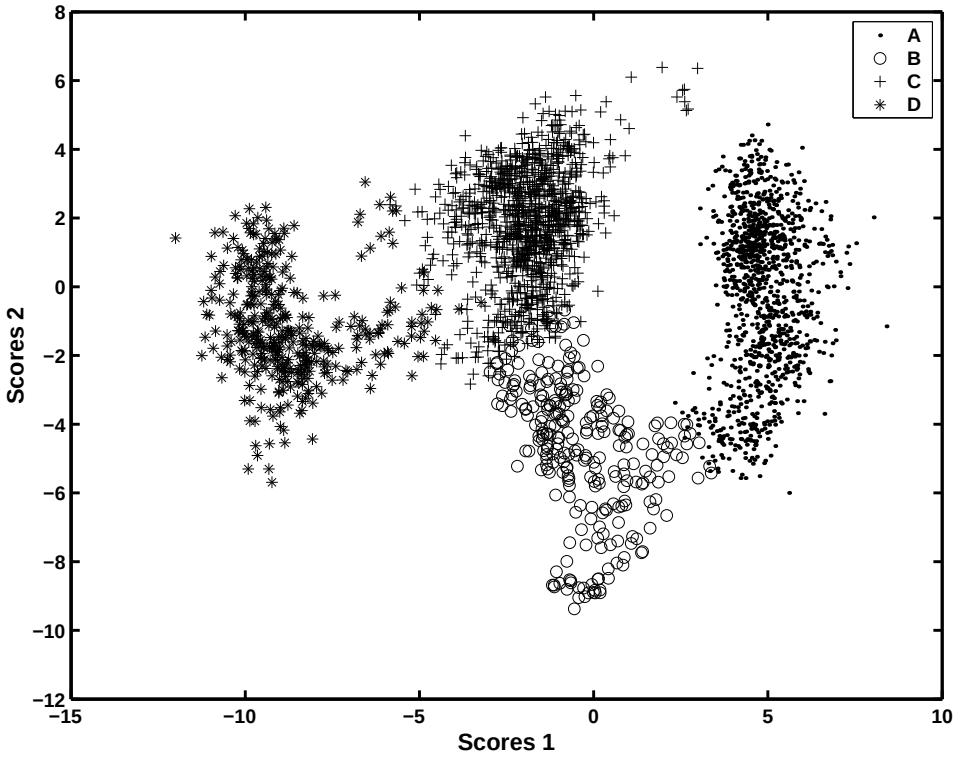


Figure 4.14: k-means classified clusters in PCA score space

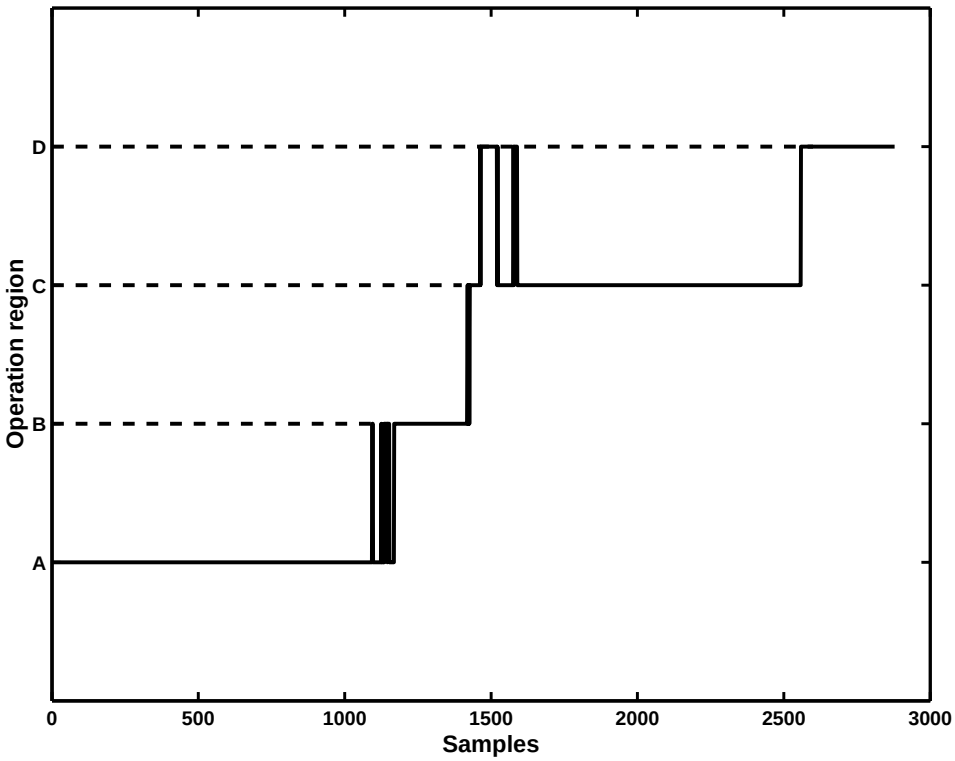


Figure 4.15: Class patterns in the polyester film process data

4.5.2 Fault visualization

After removing ambiguous samples from the transitional regions, we perform PCA based on all remaining samples to get an overall view of four classes as shown in Figure 4.16. We also perform global FDA on this reduced data set, the results are shown in Figure 4.17. By comparing Figure 4.16 and Figure 4.17 we observe that each cluster is more compact and better separated in FDA Fisher space than in PCA score space. It is interesting to note that, in FDA Fisher space, fault C is closer to normal region A than faults B and D, which is consistent with SPE and T^2 plots (Figure 4.12) where fault C has smaller SPE and T^2 values than faults B and D, while we do not observe this from PCA score plot. The comparison shows that PCA seeks directions that are efficient for representation while FDA seeks directions that are efficient for discrimination.

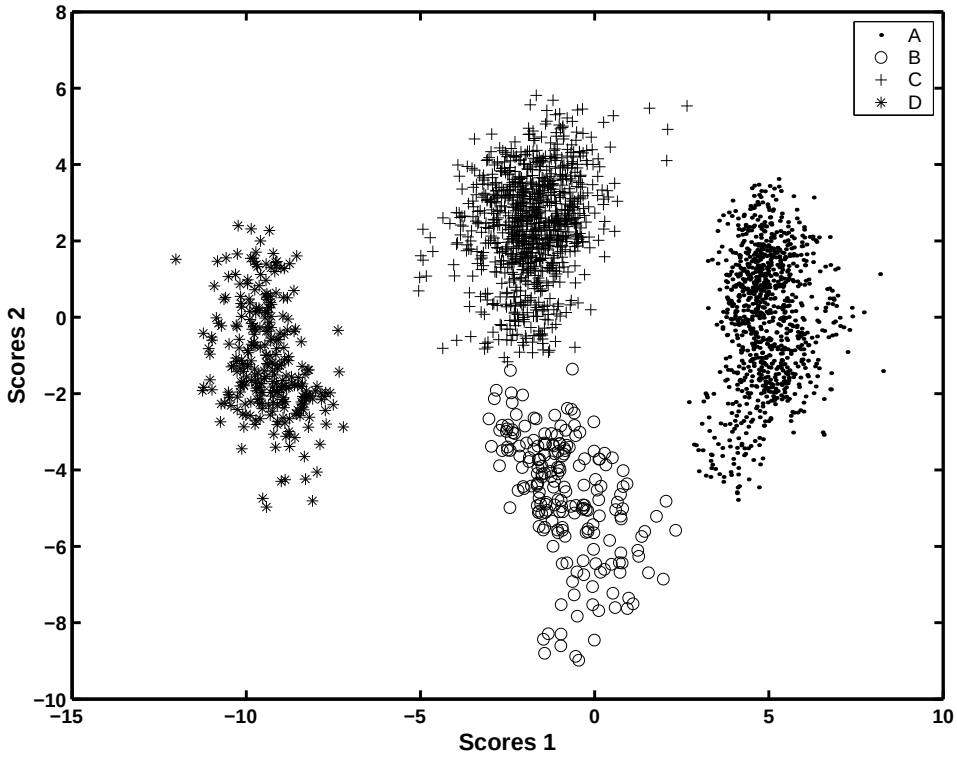


Figure 4.16: Clusters in PCA score space after deleting transitional samples

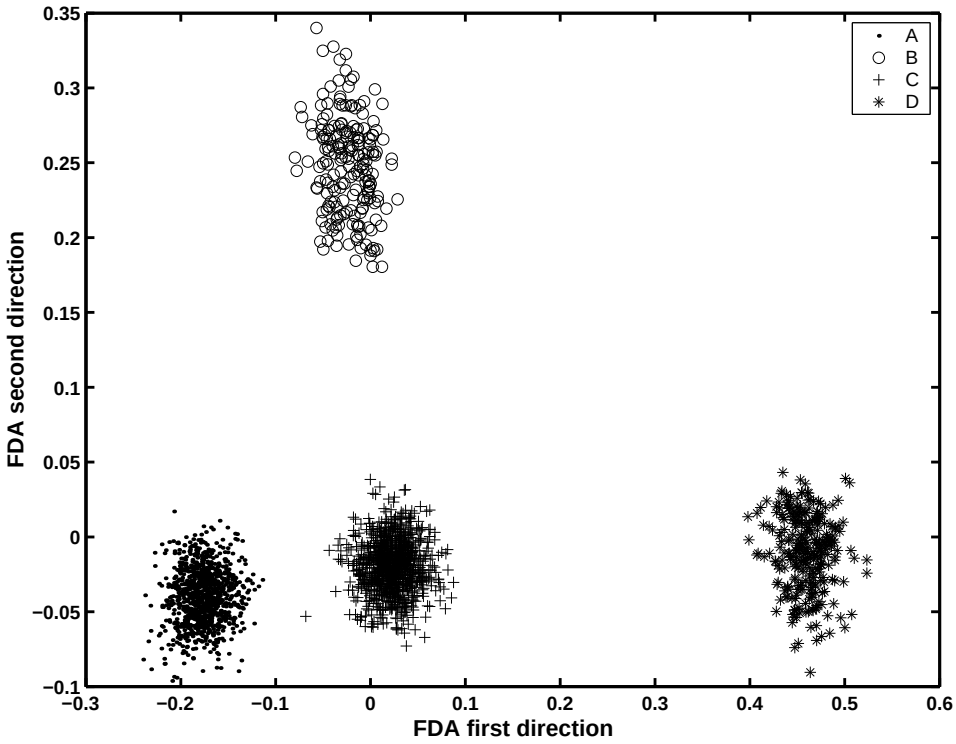


Figure 4.17: Clusters in FDA Fisher space after deleting transitional samples

4.5.3 Fault diagnosis

After the process data is classified into disjoint classes, the proposed pair-wise FDA is applied to the diagnosis of faults B, C and D. The contribution plots based on FDA fault directions for faults B, C and D are given in Figure 4.18 on left hand side (Figures (a), (c) and (e), respectively). For comparison, contribution plots based on the PCA model are also given in Figure 4.18 on right hand side (Figures (b), (d) and (f), respectively).

- **Fault B:** The contribution plot based on FDA fault direction, Figure 4.18 (a), indicates that several variables in extrusion zone contribute to this fault. The variable plots in Figure 4.19 show that oscillation of several temperature loops (Variables 25 and 28) and a step change in the power of extrusion filter (Variable 32) caused this fault. However, from the contribution plot based on PCA, Figure 4.18 (b), only Variable 28 is identified while Variables 25 and 32 are not identified.

- **Fault C:** The contribution plot based on FDA fault direction, Figure 4.18 (c), reveals that a group of variables in melt pipes zone 1 caused this fault. Among them, Variables 31 and 32 are the biggest contributors. The variable plots in Figure 4.20 show that offset in die flange powers (Variables 31 and 32) caused this fault. The contribution plot based on PCA, Figure 4.18 (d), also indicate that the fault was occurred in melt pipes zone 1 as a group of variables in that zone are identified. However, Variables 13 and 96 are also identified as two of the biggest contributors.

Since Variable 13 is located at extrusion zone while Variable 96 is located at tenter zone, it is unlikely that these two variables contribute to the fault occurred in the melt pipes zone 1. Besides, changes in Variable 96 as shown in Figure 4.20 would lead to larger SPE or T^2 value in Fault C than in Fault B, which is not true as we can see from Figure 4.12.

- **Fault D:** The contribution plot based on FDA fault direction, Figure 4.18 (e), indicates that this fault was occurred in tenter zone because several variables in that zone are identified. Process knowledge indicates that this fault is caused by the power drops in STR reheating (Variable 82) and STR crystallizer (Variables 96 and 98) and power increase in STR cooling (Variable 99). These variables together control the film temperature in tenter zone. These changes affected several other variables down-stream, but they do not have any impact before the die zone. The contribution plot based on PCA, Figure 4.18 (f), does not clearly indicate where this fault occurs and the contributing variables spread across the whole process, which is unlikely true. From variable plots in Figure 4.21, we observe sudden changes in Variables 96 and 99. We also observe that Variable 32 is not likely the root cause of this fault because there is no obvious mean or variance change during that period.

In summary, contribution plots based on FDA fault directions give better indications on which variables contribute to the faults than contribution plots based on PCA. The contribution plots based on PCA has more difficulty

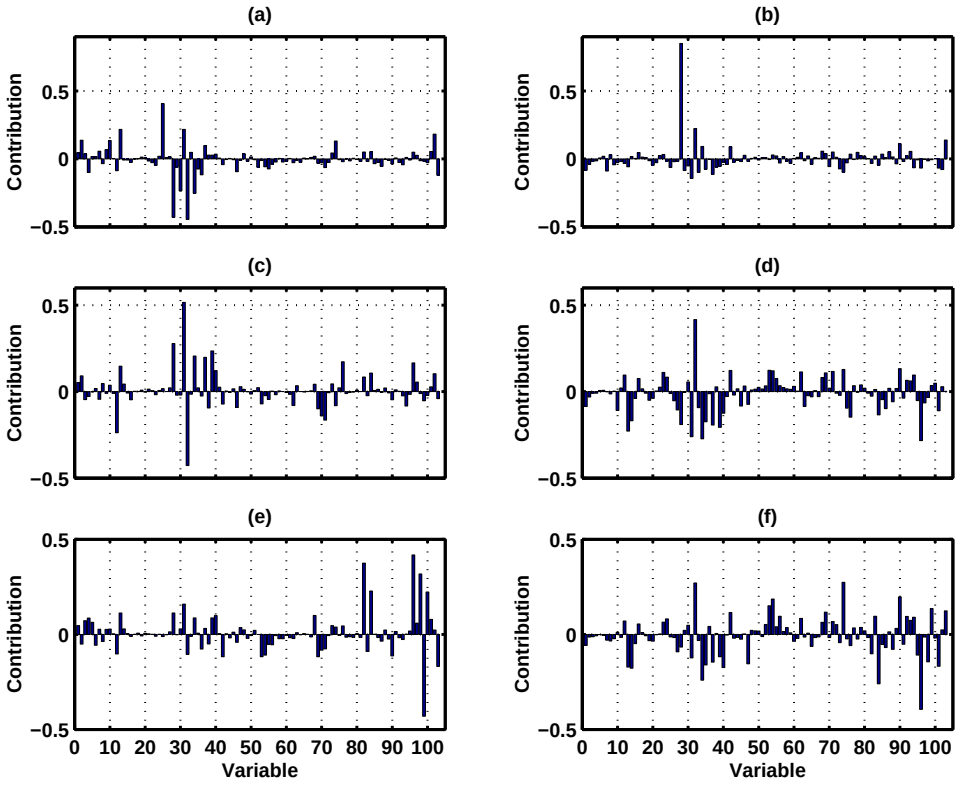


Figure 4.18: Contribution plots based on FDA and PCA

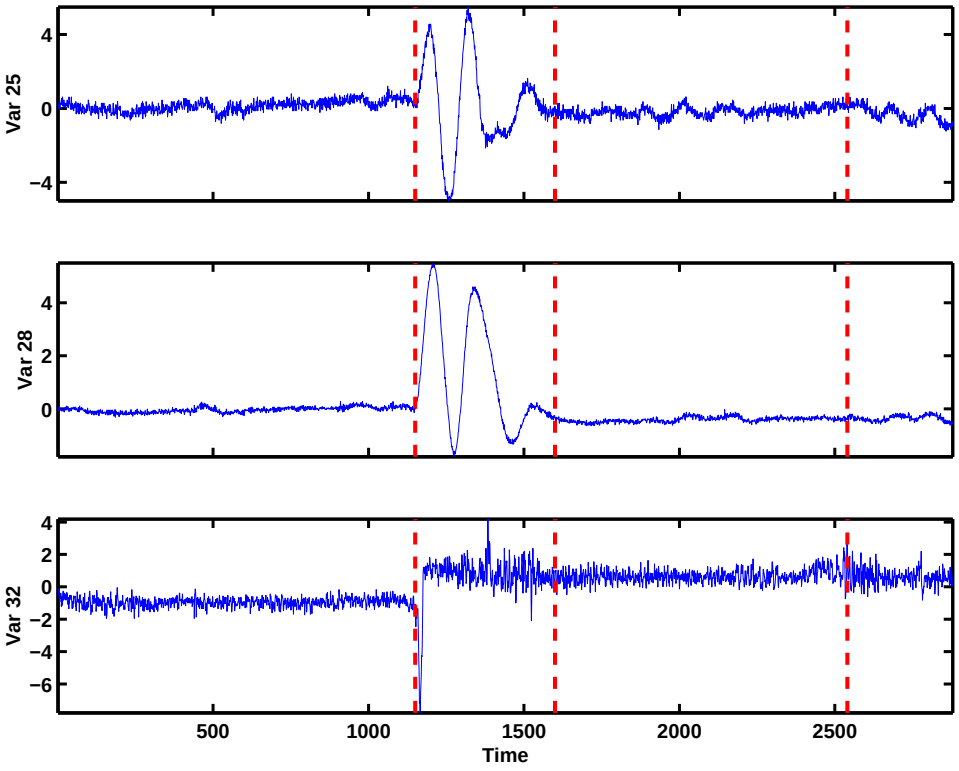


Figure 4.19: Variables 25, 28, and 32 after scaling

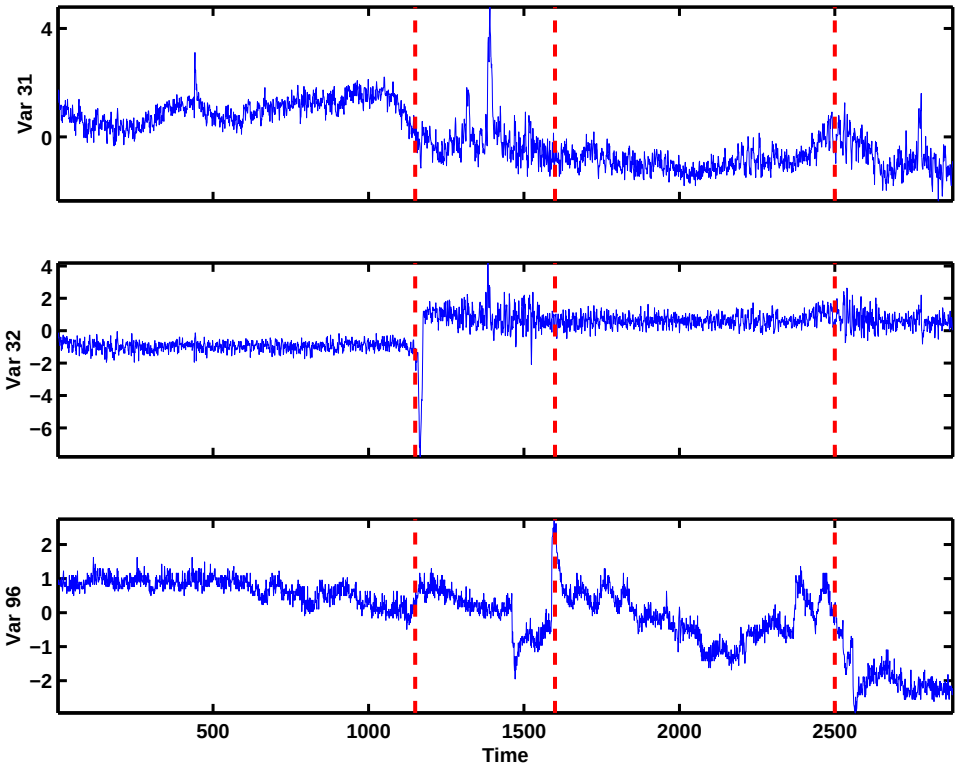


Figure 4.20: Variables 31, 32, and 96 after scaling

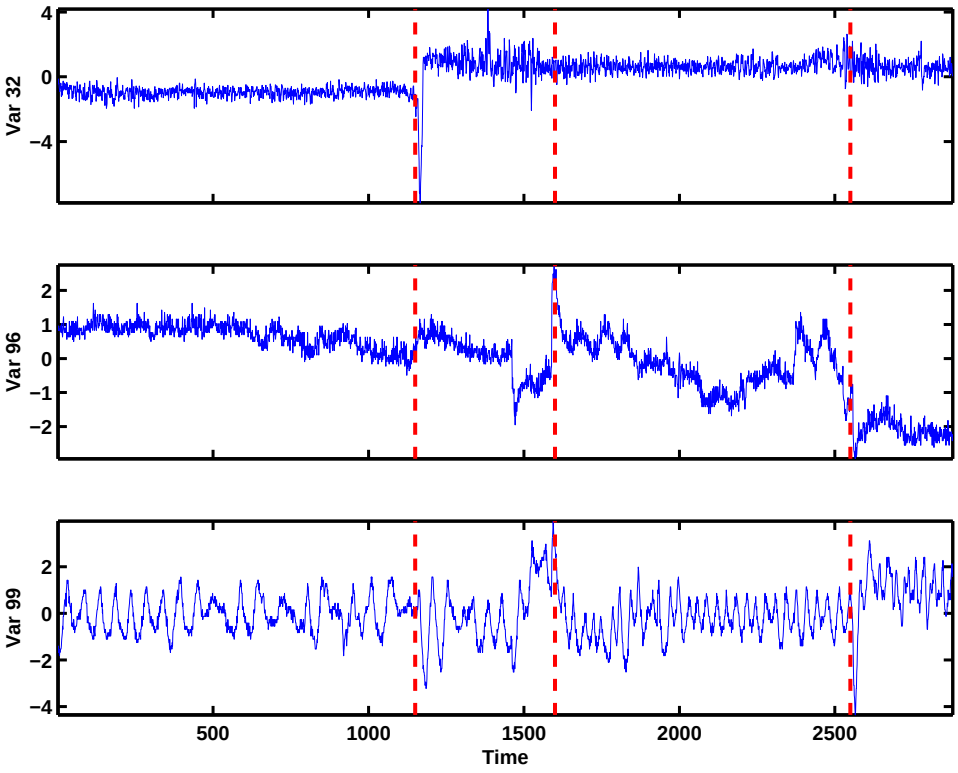


Figure 4.21: Variables 32, 96, and 99 after scaling

in fault diagnosis at the downstream of the process because of the spread effect of the fault variables, while contribution plots based on FDA works consistently across the process. The superior results of FDA fault directions are due to the fact that FDA models the fault cluster as well as the normal data cluster, while PCA models only the normal data cluster.

4.6 Conclusions

In order to use historical process data for process monitoring, it is imperative to isolate normal process data from a mixture of normal and abnormal historical data. The proposed three-step procedure described in this chapter, consisting of pre-analysis, visualization, and diagnosis, has been successfully applied to an industrial polyester film process. In the pre-analysis step for fault diagnosis, the k-means clustering method is applied in conjunction with the PCA score chart, SPE chart and T^2 chart to isolate data into different classes which correspond to different process operating regions. A clear fault visualization of high-dimensional data is obtained by applying global FDA to normal and fault data. A contribution plot based on the fault directions in pair-wise FDA is proposed to enhance the ability of fault diagnosis in multivariate statistical monitoring. The proposed method is applied to a simulated quadruple-tank process, where diagnosis of the sensor fault and leakage fault has been successfully conducted. In addition, the proposed method is applied to the industrial polyester film process and provides better fault diagnosis than PCA based contribution plots.

It is always desirable in practice to have a method that can automatically isolate historical process data into normal and fault clusters. This, however, cannot be achieved with the use of statistical methods alone. Process knowledge must be incorporated to tell the normal region from abnormal clusters and determine the total number of clusters. Once the normal data and the number of fault clusters are determined, the remaining steps proposed in this chapter are fairly automatic. Finally, the proposed method integrates PCA, FDA and clustering analysis to take advantage of the strength of each algorithm for a complete solution.

Chapter 5

Multivariate Visualization in Statistical Process Monitoring

Multivariate visualization techniques have been developed in the past three decades in the fields of statistics, artificial intelligence and computer graphics, and have been widely used in these fields as fundamental tools to visually analyze and explore multivariate data. However, these multivariate visualization techniques are typically applicable to relatively small data set which may make these techniques not applicable to the visualization of chemical processes where massive amount of data are generated on the daily basis. Besides, most of these techniques have been applied to static systems or to visualizing static properties of dynamic systems, while most chemical processes are dynamic systems and the visualization of the dynamic systems is a more difficult, largely unsolved issue. In this work, the commonly used visualization techniques applied to relatively small static systems are evaluated in the context of large dynamic systems. Dynamic parallel coordinates (DPC) is proposed to visualize large data sets and capture dynamic characteristics. Several factors which affect the quality of visualization are discussed. Variable grouping is introduced to reduce clutter in handling large data sets and hierarchical visualization scheme is proposed based on variable grouping to provide

a general framework for visualization and exploration of large multivariate data sets. Alternatively, instead of visualizing the original high dimensional raw data, some projection techniques have been developed to transfer high dimensional data into some low dimensional space. In this work, principal component analysis (PCA), partial least squares (PLS) and class-preserving projection (CPP) are evaluated for class visualization of high dimensional data. We also demonstrate that some commonly used classification methods such as Fisher discriminant analysis (FDA) and support vector machines (SVM) can be tailored for high dimension class visualization. A binary-tree approach and a cross-selection approach are proposed based on SVM. The performance of PCA, PLS, CPP, FDA and two approaches based on SVM are compared using an industrial data set. The visualization of process dynamics in the transformed space is investigated as well.

5.1 Introduction

The visualization of raw data, transformed data, and various analysis results has been important since the start of chemometrics in the late 1960s and early 1970s, and it is arguable that visualization and chemical interpretation of plots form the true nature of chemometrics [29].

Multivariate visualization techniques have been developed in the fields of statistics, artificial intelligence and computer graphics and have been widely used in these fields as fundamental tools to allow human eyes to detect special structures in data [62, 98]. In general, we are interested in the visualization

of the static properties (such as outliers, clusters and variable correlations) and dynamic properties (such as process drifts, shifts and oscillations) which can be visualized either in the original space or in the transformed space (see Figure 5.1).

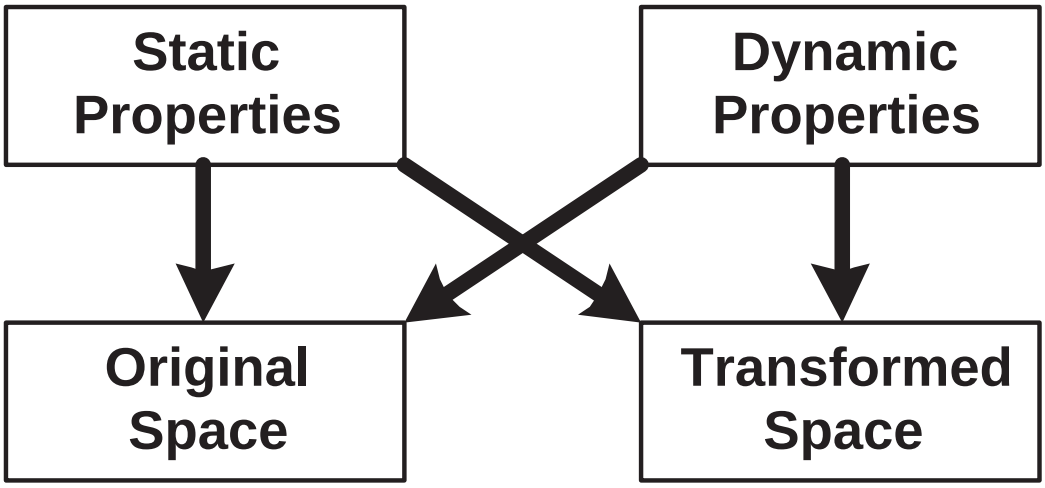


Figure 5.1: Visualization of static and dynamic properties in the original or transformed spaces

In today's chemical industry, massive amount of data are easily made available in computer controlled processes, but at the same time, the visualization of high dimensional data has been difficult. The classical scatter plots in chemometrics do not generalize readily beyond three dimensions. For this reason, alternative statistical multidimensional representations have been proposed by several authors, such as Chernoff's faces [11], star diagrams [25] or radial plots as its variant [107], and parallel coordinates [1, 98]. However, these methods are usually applicable to relatively small data sets. Due to this limitation, these methods cannot be applied to the visualization of chemical

processes due to a large number of samples and/or variables, and few visualization techniques have been studied and applied to process monitoring and control [1, 107]. Besides, most of the techniques mentioned above have been applied to static systems or to visualizing static properties of the dynamic systems where the time stamp of each sample is not an issue, while most chemical processes are dynamic systems and the visualization of the dynamic systems is a more difficult, largely unsolved issue.

In this work, the commonly used visualization techniques applied to static systems are evaluated in the context of dynamic systems. A new approach is proposed to accommodate large data sets and capture the characteristics of dynamic systems. Variable grouping is introduced to reduce clutter in handling large data sets and hierarchical visualization scheme is proposed based on variable grouping to provide a general framework for visualization and exploration of large multivariate data sets.

Alternatively, instead of visualizing the original high dimensional raw data, some projection techniques such as principal component analysis (PCA), partial least squares (PLS) and class-preserving projection (CPP) [21] have been developed to transform high dimensional data into some low dimensional space such that the original distances or similarities among observations are (nearly) preserved. The transformed data are then visualized as scatter plots in two or three dimensional space. In general, if high dimensional data can be represented in two or three dimensions, then outliers, variable correlations, and distinguishable clusters can often be discerned visually. In addition to the

projection methods we mentioned above, in this work, we also demonstrate that some commonly used classification methods such as Fisher discriminant analysis (FDA) [12, 46] and support vector machines (SVM) can be tailored for the visualization of high dimensional data. All these methods mentioned above are not well studied in the field of high-dimensional data visualization.

Through this chapter, two data sets from chemical processes will be used as example data sets for comparison of various visualization techniques. One is from the simulation of Tennessee Eastman process (TEP) [12] which consists of 980 samples with 52 variables. The other is an industrial polyester film process (PFP) data set which consists of 2808 samples with 103 variables. Unless otherwise specified, all data sets are subtracted by the mean of the normal operation data and then divided by the standard deviation of the normal operation data such that the normal operation data have zero mean and unit variance.

This chapter is organized as follows: Section 2 gives an overview of the most commonly used multivariate visualization techniques for the visualization of static properties in the original variable space. The visualization of process dynamics in the original variable space is presented in Section 3. Section 4 presents visualization of data clusters in the transformed space using some projection methods, including PCA, PLS, CPP and methods we proposed based on FDA, and SVM. Section 5 includes the discussion of the visualization of process dynamics in the transformed space, and we conclude with summary of our contributions in Section 6.

5.2 Visualization of Static Properties in the Original Variable Space

As discussed in Section 5.1, many techniques have been developed for the visualization of multivariate data sets that are becoming increasingly common. In this section, several approaches such as scatter plots and parallel coordinates are reviewed and their advantages and limitations are discussed.

5.2.1 Scatter plots

Scatter plots are one of the oldest and most commonly used methods to visualize high dimensional data in 2-D or 3-D spaces. There are two ways in which scatter plots are used to represent the data. One way is to visualize the n-dimensional raw data by generating multiple two-dimensional scatter plots for pairs of dimensions. The other way is the scatter plots based on the projection methods such as PCA, three-way analysis, FDA etc.. The later will be covered in Section 4 and here we refer to the former. An example of such scatter plots of the TEP data set for 5 variables is presented in Figure 5.2 where we observe larger variation in variable 9 than in other four variables. Advantages of scatter plots include ease of interpretation and relative insensitivity to the number of samples in the data set. One major limitation of scatter plots is that they are most effective with small numbers of variables, as increasing the dimensionality results in decreasing the screen space provided for each projection [103]. For example, the original data set we used to generate Figure 5.2 has 52 variables while only 5 are displayed. It would be difficult

to put 52 by 52 figures into a single plot. Another limitation of scatter plots is that they are generally restricted to orthogonal views and difficult to discover relationships which span more than two dimensions. Also, it is impossible to discover the dynamic properties of the data set by inspecting scatter plots.

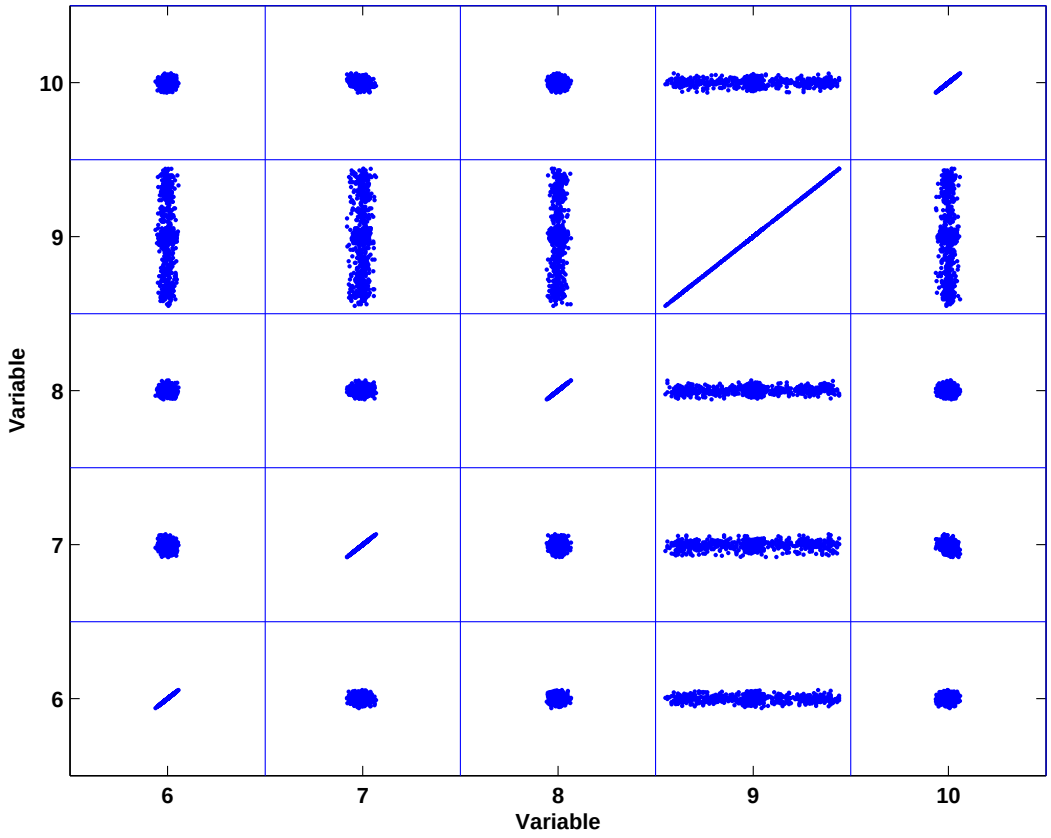


Figure 5.2: Scatter plots of TEP data with 5 variables only

5.2.2 Parallel coordinates

Parallel coordinates is a technique pioneered in the 1970's which has been applied to the visualization of a diverse set of multidimensional prob-

lems [58]. Recently it has been successfully applied to the detection of abnormal events for a waste-water treatment plant [1]. In this method, each dimension corresponds to an axis, and the axes are organized as uniformly spaced vertical lines. Each observation in n-dimensional space manifests itself as a connected set of points, one on each axis. Some properties and statistical interpretations as a projective transformation are discussed in [98]. Figure 5.3 shows an example of the parallel coordinates using the same data as in Figure 5.2 and large variation in variable 9 is detected. The advantage of parallel coordinates is that each sample is represented in a planar diagram, so each sample component has essentially the same representation [98]. Also, the individual parallel coordinate axes represent one-dimensional projections of the data and thus separation on any one axis represents a view of the data that allows the detection of clustering. The major limitation of the parallel coordinates is that large data sets can cause difficulty in interpretation; as each observation generates a line, lots of observations can lead to rapid clutter. Another limitation is that relationships between non-adjacent coordinates are much more difficult to perceive than between adjacent coordinates. Also, as the number of dimensions increases, the axes get closer to each other which makes it more difficult to perceive structure or clusters. As in scatter plots, the dynamic behavior of the data set is not revealed in parallel coordinates.

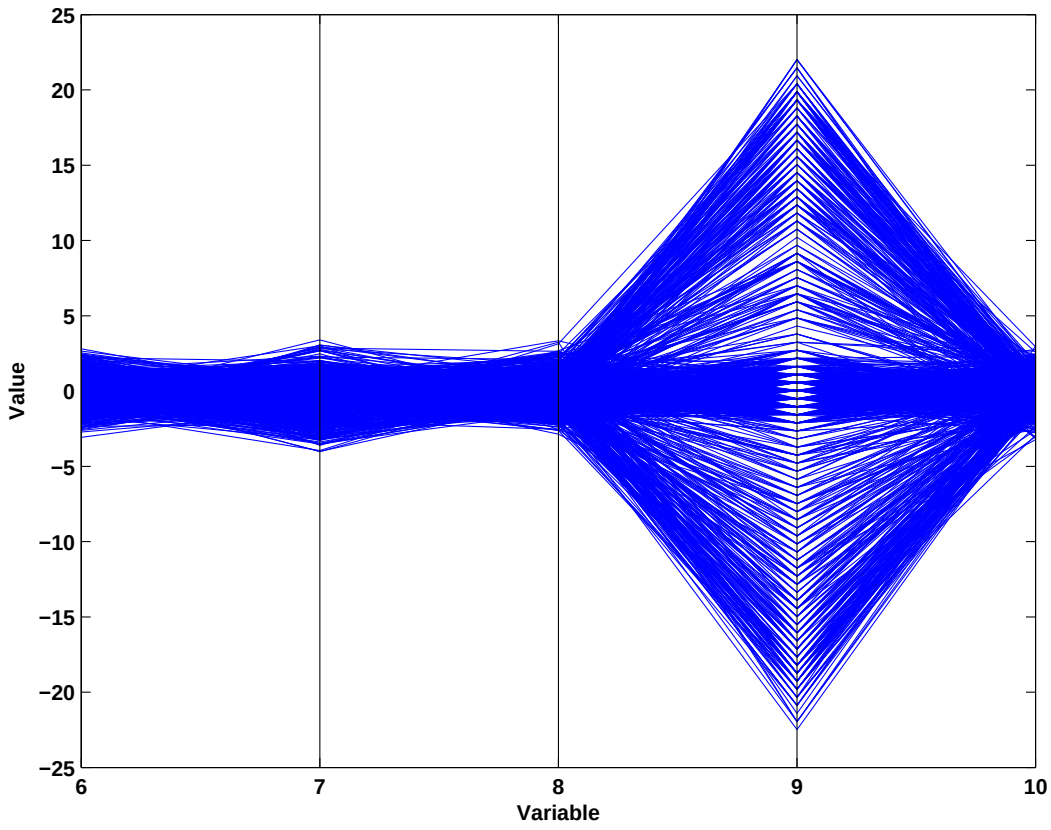


Figure 5.3: Parallel coordinates of the same data set as in Figure 5.2

5.2.3 Other types of plots

Chernoff's faces, star plots or radial plots and their variants are essentially icon or symbol based representations. Radial plots have been used in the context of business process monitoring and control where a statistical process control (SPC) chart using radial plots is used to relate the variation of the process to other variables that are being observed simultaneously with the variable that is charted [107]. This type of display allows the process to

be analyzed as a whole, while visualizing the effects from multiple process variables. However, these plots can hardly be extended to several dozens of variables which is very common in chemical processes. Also, the clearness of these plots degrades quickly as the number of samples increases.

5.3 Visualization of Process Dynamics in the Original Variable Space

In Section 5.2 we discussed the visualization of static properties in the original variable space. In this section, we focus on the visualization of the process dynamics in the original variable space. Extruded parallel coordinates will be discussed first, then an extension of extruded parallel coordinates is proposed to improve the effectiveness of the technique. Contour plots are also explored and several factors which affect the quality of visualization are discussed. Then, hierarchical visualization based on variable grouping is proposed to tackle the clutter problem and provide a general framework for handling large data sets.

5.3.1 Extruded parallel coordinates (EPC)

Displaying the variable trajectories is an important task to allow direct global visualization of the behavior of a dynamic system [34]. Traditional parallel coordinates are less effective when the dimension gets bigger. Also, traditional parallel coordinates do not reveal the system dynamic behavior. In order to visualize the behavior of higher dimensional dynamic systems, extruded parallel coordinates have been developed based on the traditional parallel coordinates [34]. Instead of using the same coordinate system for each sample, now the extruded parallel coordinates are moving along the time axis. Figure 5.4 shows extruded parallel coordinates of the TEP data set with all 52 variables. Increasing variation in several variables at the second half of the

process are uncovered. However, because both the number of samples and the number of variables are large, lines are crowded.

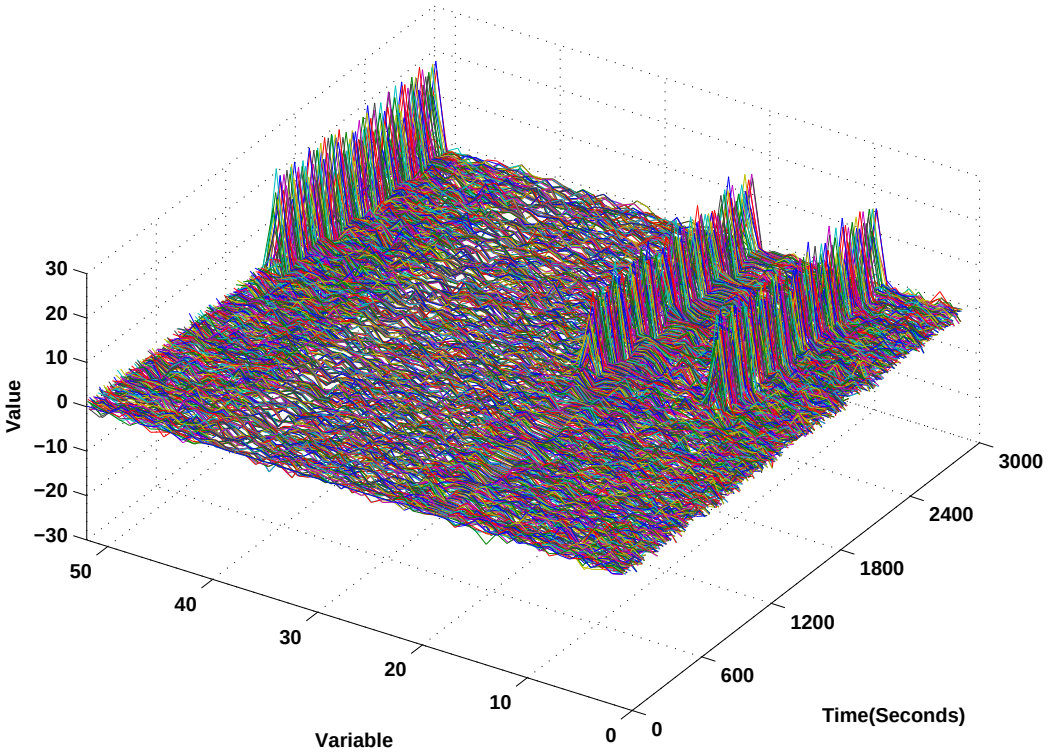


Figure 5.4: EPC plot of the TEP data

5.3.2 Dynamic parallel coordinates (DPC)

Based on the fact that the adjacent variables in EPC may not be well correlated, we propose a new parallel coordinates which eliminate the connections between adjacent dimensions; instead, we connect samples of one variable at different time together. Since this type of parallel coordinates reveals the process dynamics along the time axis, we name it dynamic parallel coordinates

(DPC). Figure 5.5 shows the same data set in dynamic parallel coordinates and it clearly reveals the dynamic behavior of the system which we do not see from the traditional parallel coordinates or difficult to see from the extruded parallel coordinates.

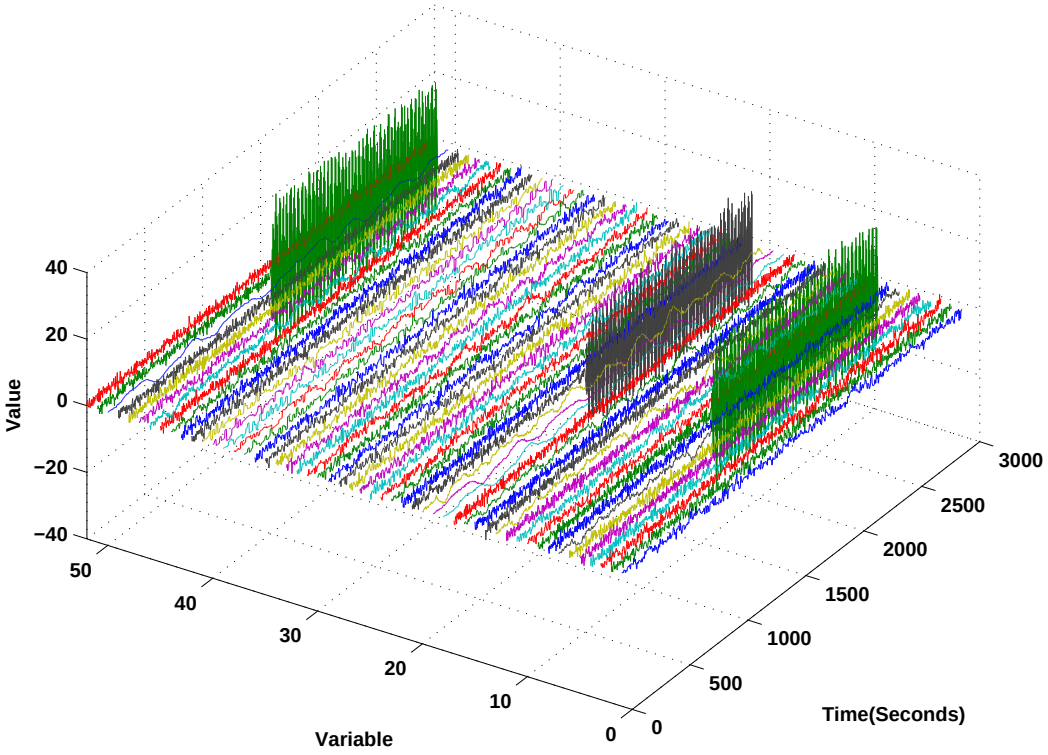


Figure 5.5: DPC plot of the TEP data

5.3.3 Contour plots

Contour plots have been widely used by geographers and the most popular classic way of visualizing peaks and valleys is using contours plots, as it is done on topographic maps. However, contour plots are not widely

used in chemometrics. There are two kinds of contour plots, 2-D contour plots in which contours are drawn on a plane, and 3-D plots that produces level curves in a 3-D space. Figures 5.6 and 5.7 show the TEP data in 2-D and 3-D contour plots. Variation increase in three variables are revealed in both figures. However, one drawback of contour plots is that they are computationally intensive to generate and spatially expensive to save if many contour lines are desired for large data sets. Another drawback is that unlike topographic maps where adjacent values in both x and y directions are spatially related, adjacent variables in Figures 5.6 and 5.7 may not be correlated very well.

5.3.4 Factors affecting visualization quality

There are several factors which play important roles in the visualization. Here we focus on scaling, smoothing and key variable identification. Scaling is one of the most important factors which affect the quality of visualization. Choosing the right scaling is as important as choosing the right visualization technique. For example, auto-scale, which scale the variable to zero mean and unit variance, is not a good way to scale the data for visualization although it is effective in other analyses such as PCA. Figure 5.8 shows the same data set as in Figure 5.5 with auto-scale where the process dynamics becomes much less observable. One effective way of scaling for visualization is to scale the data based on the process specification where the observed value is subtracted by the nominal value and divided by the range of upper and lower control limits.

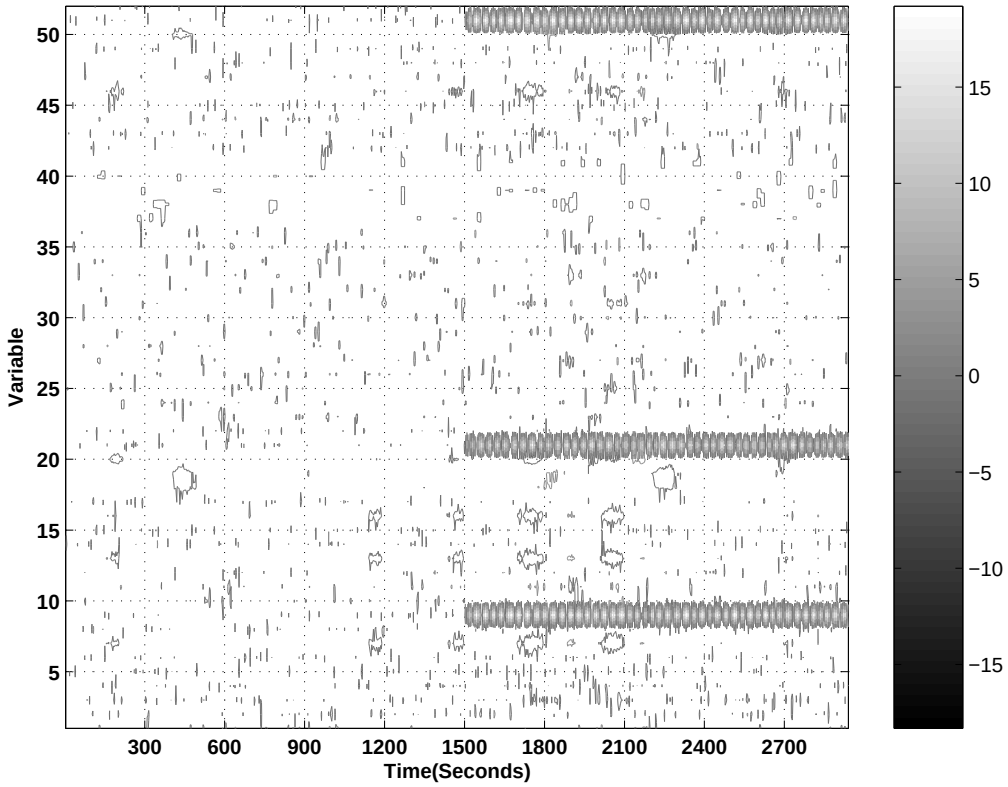


Figure 5.6: 2-D contour plot of the TEP data

If process specification is not available, which is the case for the examples used in this work, data sets are subtracted by the mean of the normal operation data and then divided by the standard deviation of the normal operation data.

Besides scaling, smoothing can be important in some cases. It can improve the quality of visualization if significant level of noise is present. Most of the low-pass filters can serve this purpose.

For large data sets, key variable identification which filters out variables

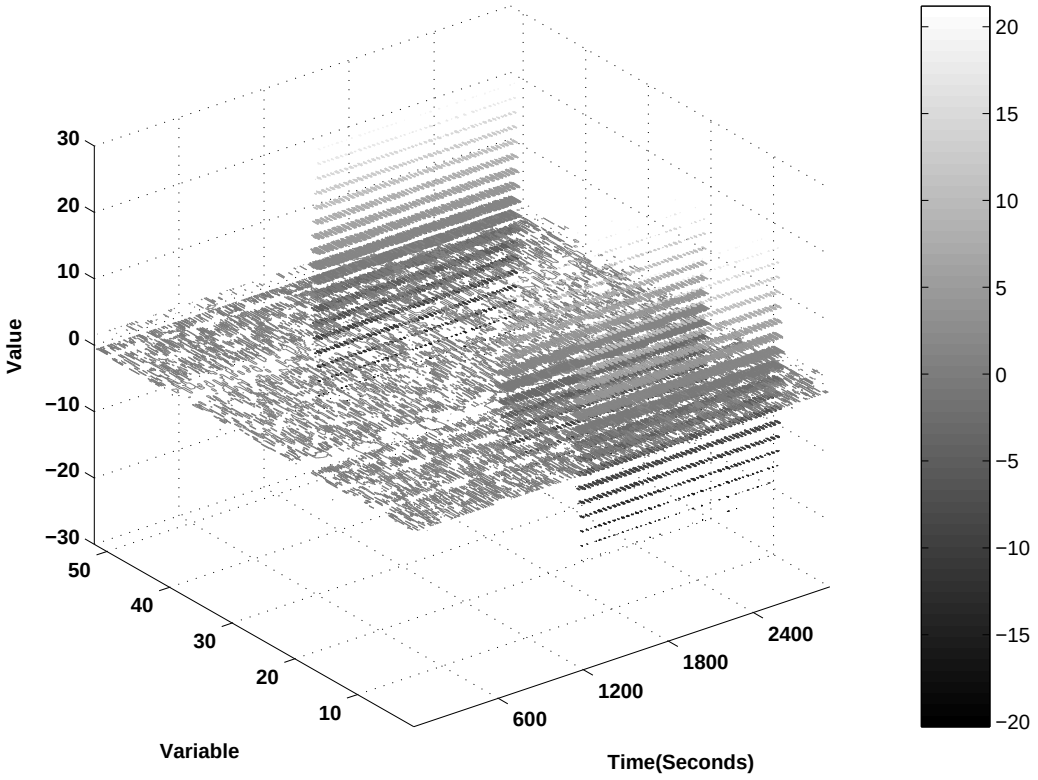


Figure 5.7: 3-D contour plot of the TEP data

with unnoticeable changes will significantly improve the clearness of the visualization. The criterion for key variable identification could be the threshold of variance or the range so that only variables with significant variance or range changes will be displayed. For example, most of the variables in Figure 5.5 experience almost no change across the process. Since we are more interested in the variables which deviate from normal operation, a key variable identification is desired to narrow down the visualizing variables. Figure 5.9 shows the key variables after we filter out variables with variances within two times

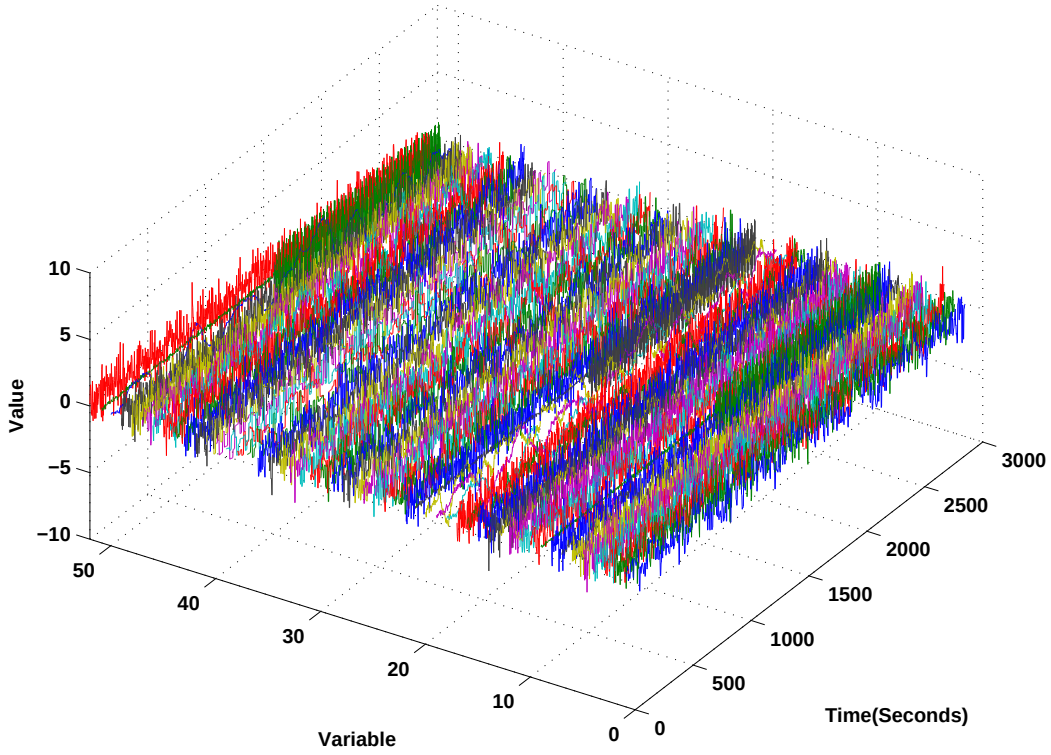


Figure 5.8: DPC plot of the TEP data with auto-scale

of their normal values.

5.3.5 Variable grouping and hierarchical visualization

To tackle the clutter problem faced by multivariate visualization techniques when analyzing large-scale data sets, in addition to key variable identification, we propose a general framework of hierarchical visualization based on variable grouping. The underlying principle of this approach is to provide multi-resolution view of the data via variable grouping. Interested region can then be explored at different levels of detail.

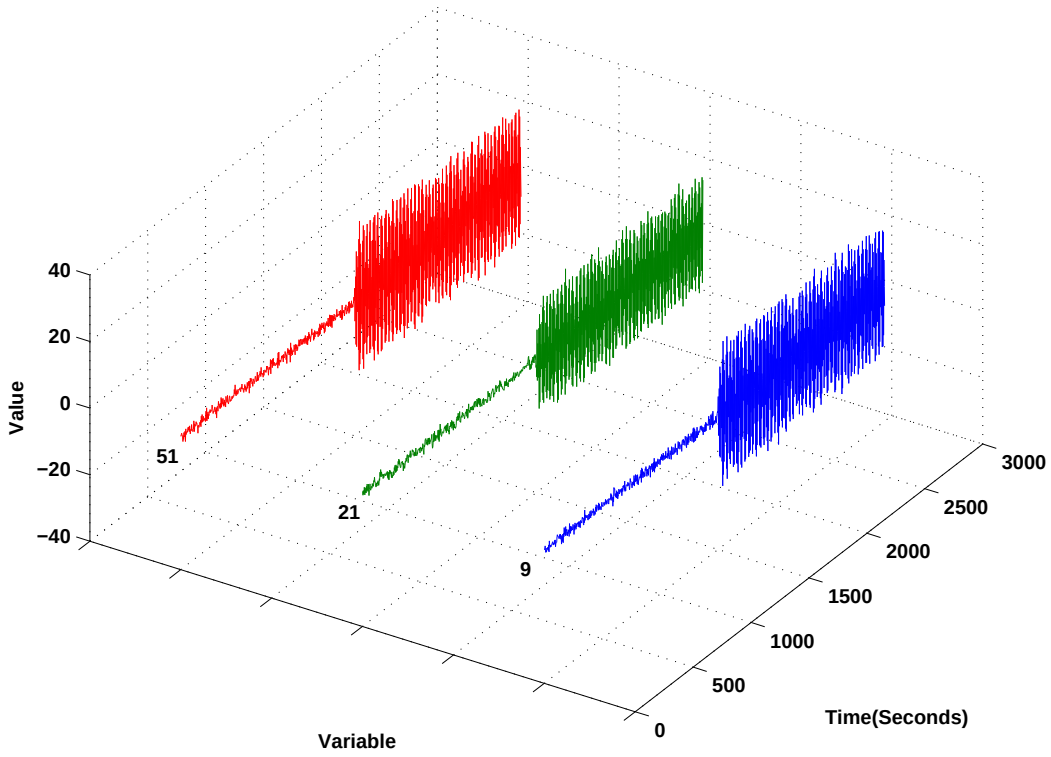


Figure 5.9: DPC plot of the TEP data with key variable identification

For chemical processes, variables can be conveniently grouped by operation unit, such as reactor, separator, or by variable type, such as temperature, pressure, flow rate. Hierarchical visualization is implemented in the following way as schemed in Figure 5.10: All variables are grouped by operation unit first; then variables associated with the same operation unit are grouped by variable type. Once a special event was detected in one unit, variables in that unit are explored by type to further hunt down the problem. In order to see the detail, individual variables in one unit with the same type could be displayed. Depending on the application, the order of grouping can be switched and one

of the grouping steps can be skipped. If necessary, key variable identification can be incorporated with any step to further reduce the number of variables to display.

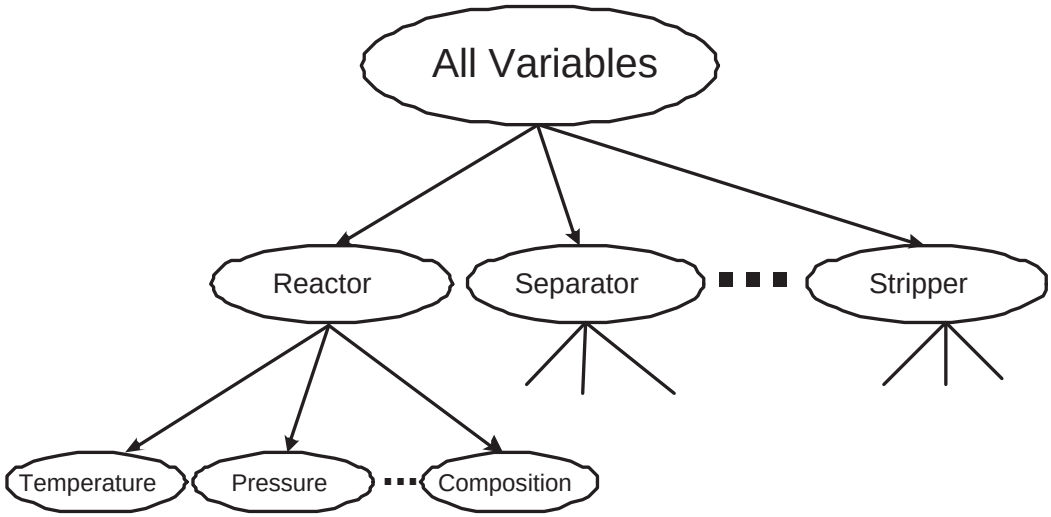


Figure 5.10: Schematic diagram of hierarchical visualization based on variable grouping

Figure 5.11 shows the DPC plot of the same data as in Figure 5.5 but with variables grouped by 4 operation units. The number in the parenthesis denotes how many variables are grouped in that unit. Variables in the same group are drawn overlapped to each other. Figure 5.11 clearly indicates that variables in all groups but the reactor are normal. Next step is to look into variables in the reactor group. Figure 5.12 shows reactor group variables grouped again by type and the number of variables in each type is in the parenthesis. While the level, pressure and compositions look normal, increasing variations in temperatures and flow rates are observed. DPC plots of individual variable

in these two groups are give in Figures 5.13 and 5.14 with abnormal variables labelled. Because reactor cooling water flow rate directly affects reactor temperature and cooling water outlet temperature, one reasonable guess of cause would be the failure of cooling water flow control, which make the truth - the fluctuation of cooling water flow rate caused by a sticking valve, very easy to understand.

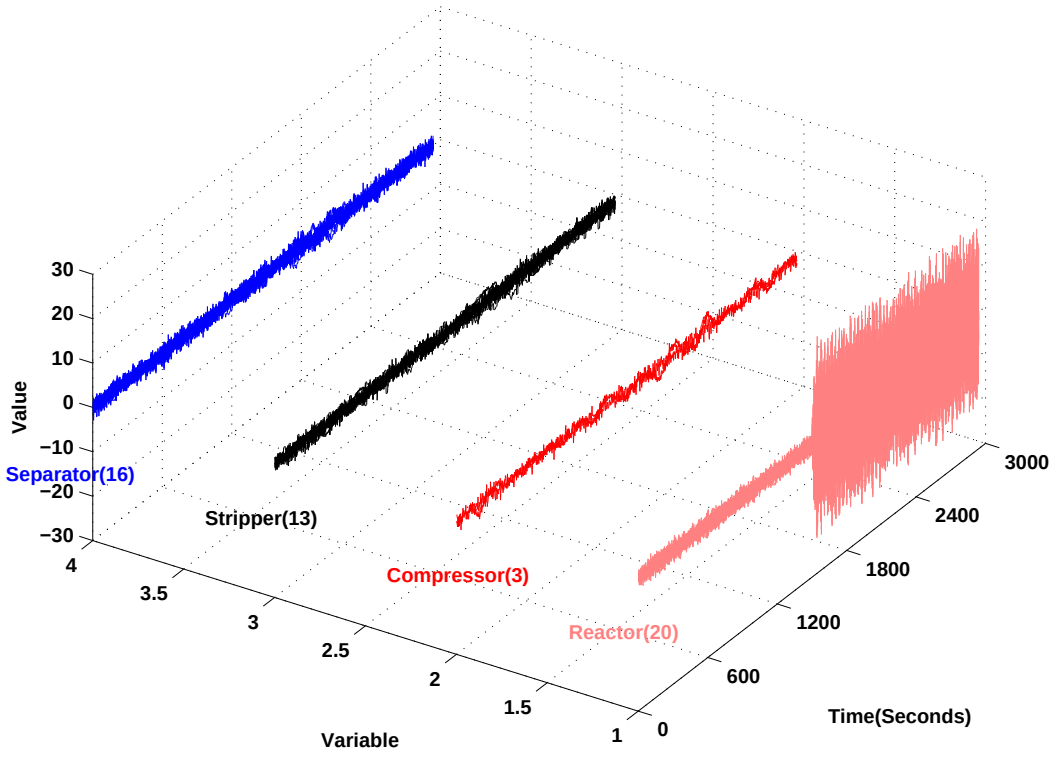


Figure 5.11: DPC plot of the TEP data with variable grouping by operation unit

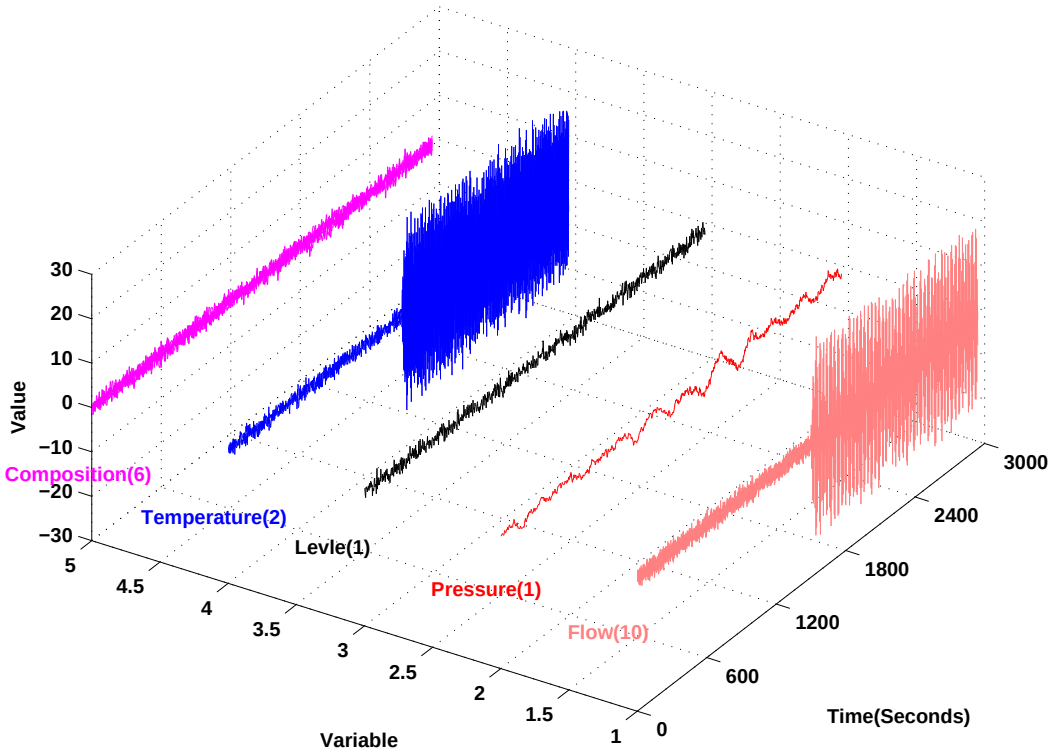


Figure 5.12: DPC plot of reactor group with variable grouping by type

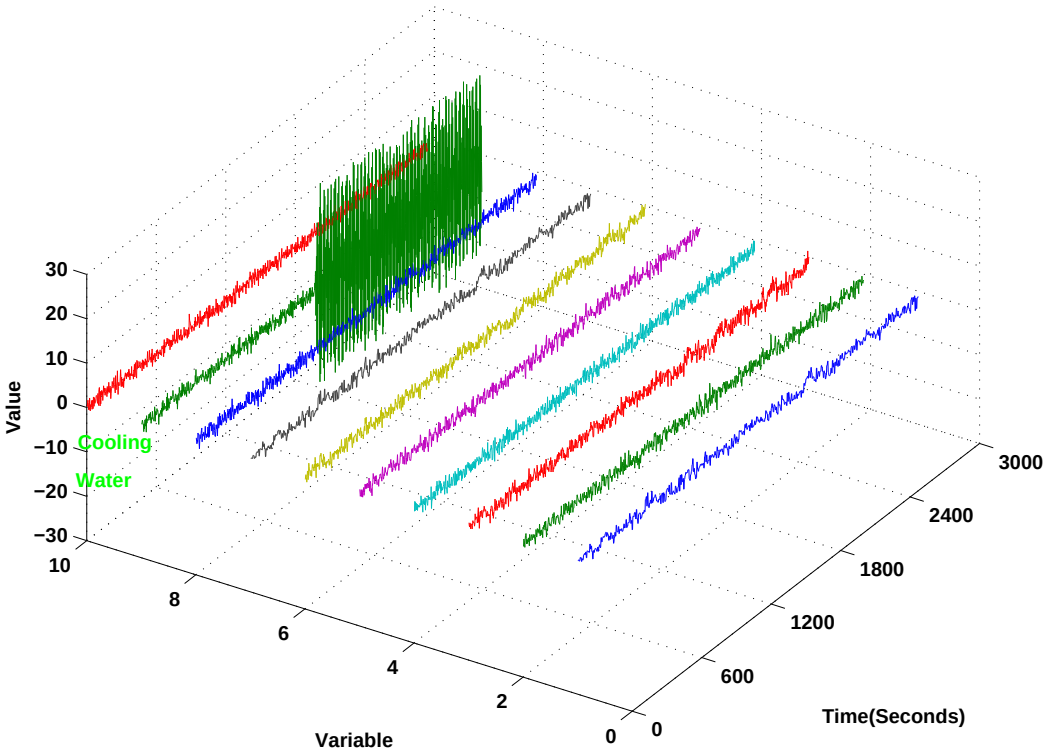


Figure 5.13: DPC plot of flow rates in reactor group

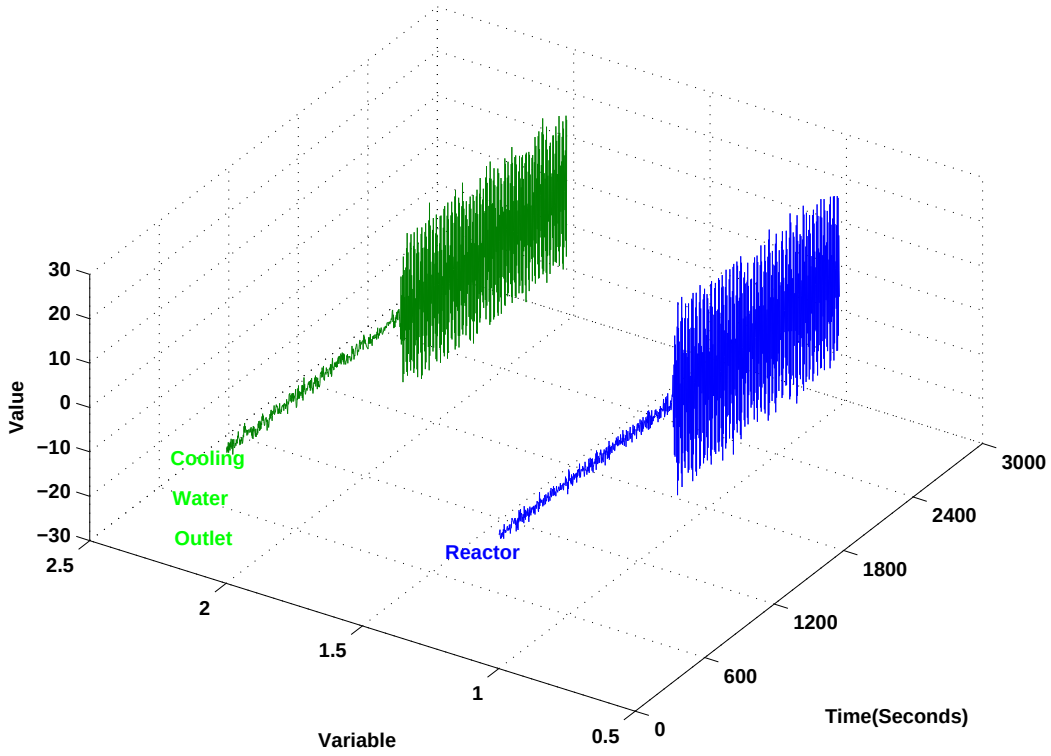


Figure 5.14: DPC plot of temperatures in reactor group

5.4 Visualization of Static Properties in the Transformed Space

In previous section, visualization of static properties in the original variable space is discussed. As we mentioned in Section 5.1, by some projection methods, special structures in the high-dimensional space, such as outliers, clusters, etc., can often be visualized in the transformed lower dimensional space. And more often than not, the process change may occur not in any single dimension but in the combination of multiple dimensions, the outliers or clusters in the data set may be better visualized in some transformed space than in the original space. In this section, we will be focusing on the visualization of data clusters and the comparison of several projection approaches in terms of visual class discrimination. The terms class and cluster are used exchangeable in this work. The polyester film process (PFP) data is used in this section to compare the performance of different visualization techniques and it is known that the data set consists of four classes – one normal operation and three different types of faults, and class label is assigned to each sample. Two most commonly used projection method – PCA and PLS, are reviewed and applied to the PFP data first. Then two commonly used classification methods – FDA and SVM, are reviewed and tailored to the needs of visualization. CPP shares some common characteristics with FDA, so it is reviewed after FDA and compared to other methods.

5.4.1 Principal component analysis (PCA)

PCA can significantly reduce the dimensionality of the data set by transforming a number of correlated variables into a much smaller number of uncorrelated variables called principal components (PC's) [23]. The first PC captures as much of the variability in the data as possible, and each succeeding component captures as much of the remaining variability as possible. It has been proven that the representation given by PCA is an optimal linear dimension reduction technique in the mean-squared sense [63]. Let $X^0 \in \Re^{n \times m}$ denote the raw data matrix with n samples and m variables. It is first scaled to a matrix X with zero mean for covariance-based PCA and, with unit variance for correlation-based PCA. By either the NIPALS [102] or a singular value decomposition (SVD) algorithm, the scaled matrix X is decomposed as follows:

$$X = TP^T + \tilde{X} \tag{5.1}$$

where $T \in \Re^{n \times l}$ and $P \in \Re^{m \times l}$ are the score matrix and the loading matrix, respectively, and $\tilde{X}$ is the residual matrix. The PCA projection reduces the original set of m variables to l PC's where $T = XP$ contains the projections of the observations or so called transformed observations. By drawing the first several score vectors in 2-D or 3-D scatter plot, the maximum possible variation captured by first several principal components is presented visually. Notice that PCA does not make use of the class information.

Since PCA is good at representing the data, if the PCA model is built based on data from all classes, the first few PCs would capture the most

significant common characteristics of samples among different classes. So the best separation of samples from different classes may not be observed within the first few PC spaces. For example, 2-D scatter plot of the first two PCs in Figure 5.15 and 3-D scatter plot of the first three PCs in Figure 5.16 do not separate 4 classes in the PFP data very well.

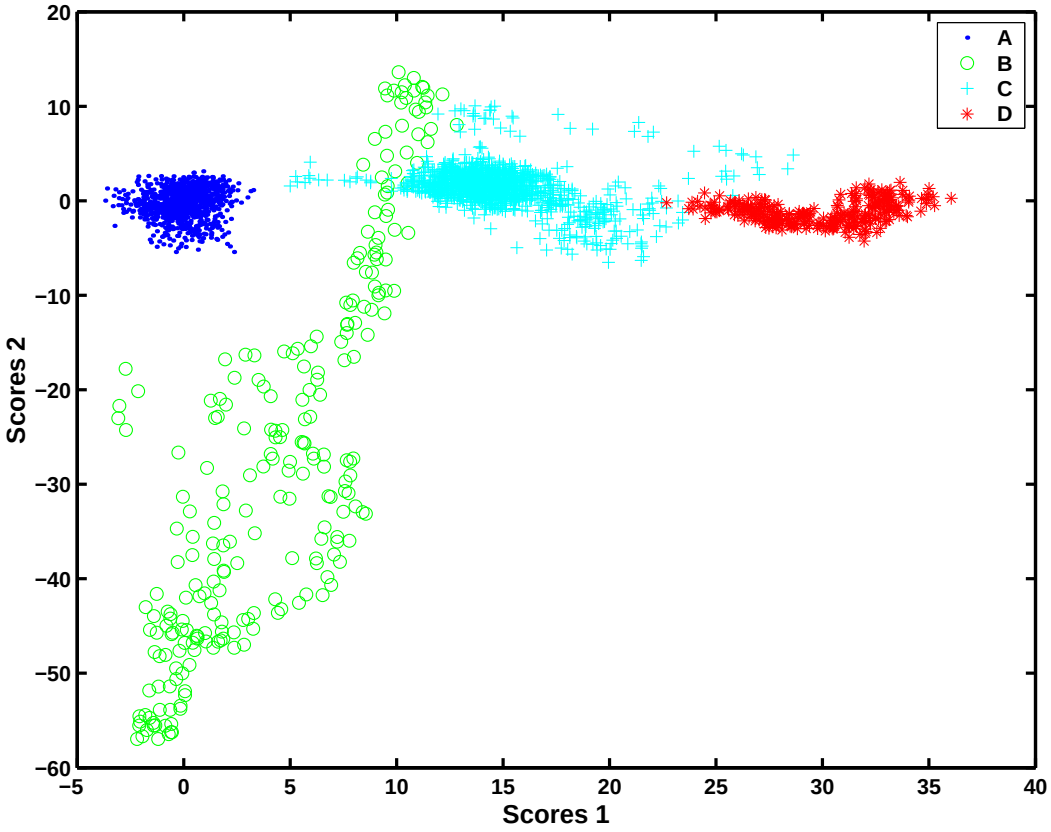


Figure 5.15: 2-D PCA score plot of the PFP data

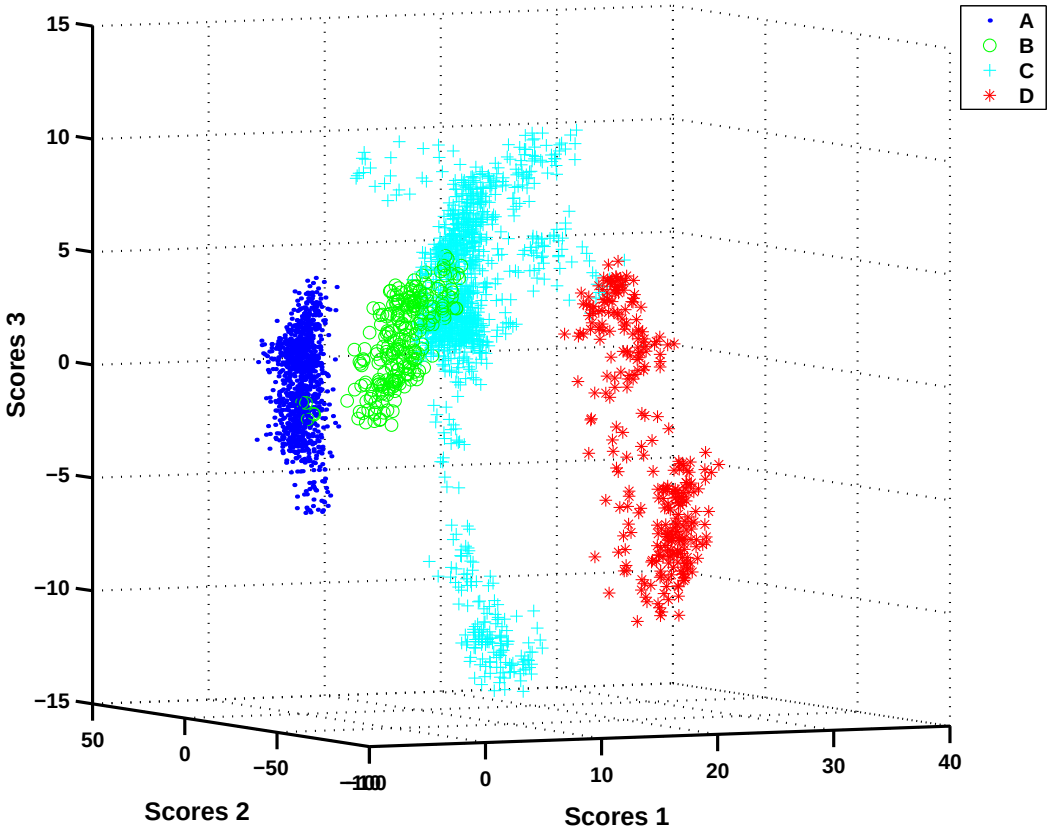


Figure 5.16: 3-D PCA score plot of the PFP data

5.4.2 Partial least squares (PLS)

PLS, also known as Projection to Latent Structures, is a projection method which maximize the covariance between the predictor (independent) matrix X and the predicted (dependent) matrix Y for each component of the reduced space [13, 28]. For the visualization purpose, $Y \in \Re^{n\times 1}$ is designed to contain the class label for each observation. This type of PLS sometimes is referred as discriminant PLS. As in PCA, X and Y are mean-centered and properly scaled, and then decomposed into the same form as that in PCA. X

is decomposed into a score matrix $T \in \Re^{n \times l}$ and a loading matrix $P \in \Re^{m \times l}$, plus a residual matrix $\tilde{X} \in \Re^{n \times m}$:

$$X = TP^T + \tilde{X} \tag{5.2}$$

Similarly, Y is decomposed into a score matrix $U \in \Re^{n \times l}$ and a loading matrix $Q \in \Re^{1 \times l}$, plus a residual matrix $\tilde{Y'} \in \Re^{n \times 1}$:

$$Y = UQ^T + \tilde{Y'} \tag{5.3}$$

PLS determines the loading and score vectors which are correlated with Y while describing a large amount of the variation in X by regressing U to T:

$$\hat{U} = TB \tag{5.4}$$

where $B \in \Re^{l \times l}$ is the diagonal regression matrix, which gives

$$Y = TBQ^T + \tilde{Y} \tag{5.5}$$

where $\tilde{Y}$ is the prediction error matrix. Since PLS takes the class information into account when build the model, theoretically, it should perform better than PCA in terms of class discrimination. 2-D and 3-D PLS score plots in Figures 5.17 and 5.18 show improvement over PCA score plots.

5.4.3 Fisher discriminant analysis (FDA)

FDA is a widely used technique in pattern classification. The basic idea of FDA is to find the Fisher optimal discriminant vector such that the ratio of

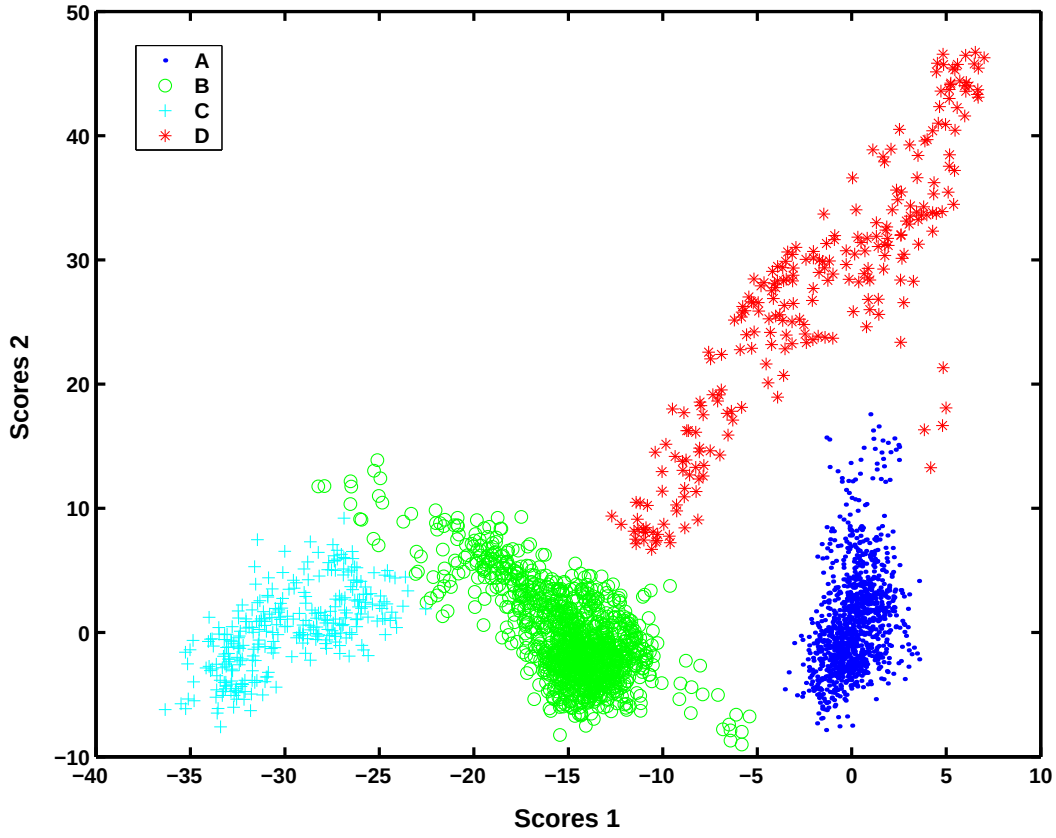


Figure 5.17: 2-D PLS score plot of the PFP data

the between-class scatter to the within-class scatter is maximized. The higher-dimensional feature space then can be projected onto the obtained optimal discriminant vectors for constructing a lower-dimensional feature space. Let $X \in \Re^{n \times m}$ be a set of m-dimensional samples $x \in \Re^m$ and the matrix X_i is the subset containing n_i rows of X corresponding to the samples from class i. If $\bar{x}_i$ is the m-dimensional sample mean for class i given by

$$\bar{x}_i = \frac{1}{n_i} \sum_{x \in X_i} x \tag{5.6}$$

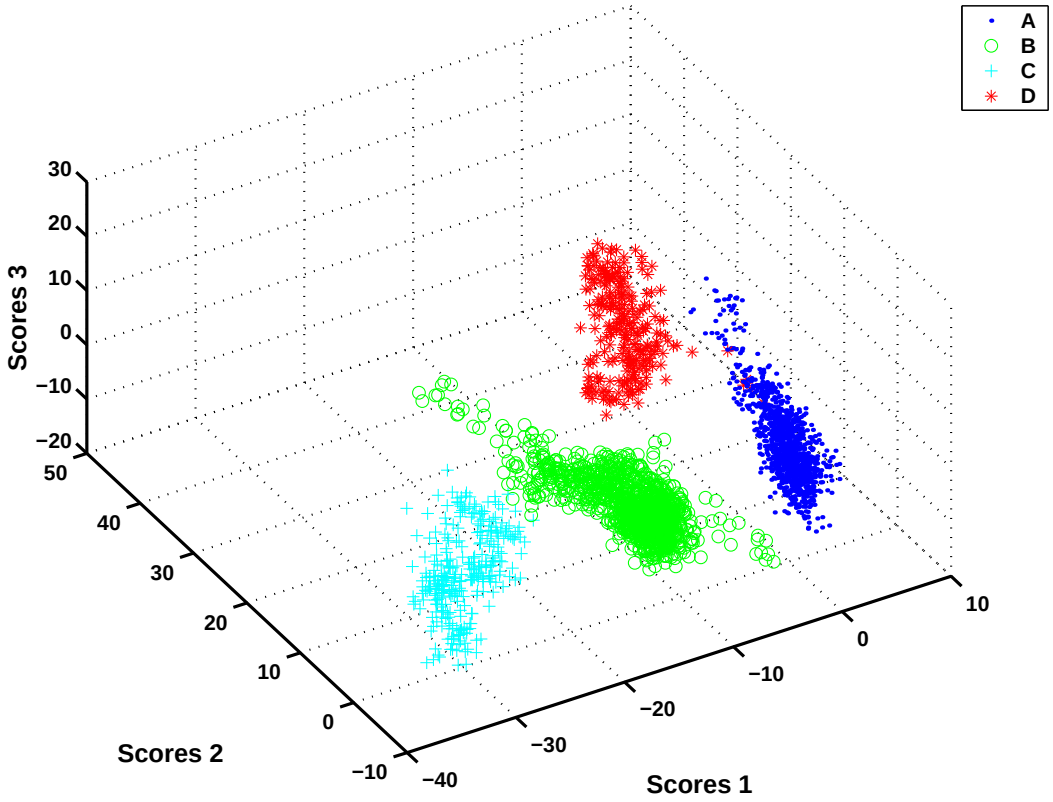


Figure 5.18: 3-D PLS score plot of the PFP data

then the within-class scatter matrix is given by

$$S_w = \sum_{i=1}^{c} P(\omega_i) S_i \tag{5.7}$$

where

$$S_i = \frac{1}{n_i} \sum_{x \in X_i} (x - \bar{x}_i)(x - \bar{x}_i)^T \tag{5.8}$$

is the within-class scatter matrix for class i and $P(\omega_i)$ is the a priori probability of class i, generally, $P(\omega_i) = 1/c$.

Let $\bar{x}$ be the mean vector of all samples in X, the between-class scatter

matrix is defined by

$$S_b = \sum_{i=1}^{c} P(\omega_i)(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T \tag{5.9}$$

The optimal discriminant direction is found by maximizing the Fisher criterion:

$$J(\varphi) = \frac{\varphi^T S_b \varphi}{\varphi^T S_w \varphi} \tag{5.10}$$

where the maximizer φ is the Fisher optimal discriminant direction which maximizes the ratio of the between-class scatter to the within-class scatter. It is easy to show that a vector φ that maximize $J(\cdot)$ must satisfy

$$S_b \varphi = \lambda S_w \varphi \tag{5.11}$$

for some constant λ, which is a generalized eigenvalue problem. If S_w is non-singular, we can obtain a conventional eigenvalue problem by writing

$$S_w^{-1} S_b \varphi = \lambda \varphi \tag{5.12}$$

The second Fisher direction is computed to maximize the same ratio among all directions perpendicular to the first Fisher direction, and so on for the remaining Fisher directions. The score matrix T is obtained by projecting the observations X onto the Fisher directions ϕ:

$$T = X\phi \tag{5.13}$$

where $\phi = [\varphi_1 \varphi_2 \cdots \varphi_l]$.

Unlike PCA which finds directions optimally representing the data set, FDA finds the optimal directions which optimally separate samples from different classes. Because of the discrimination nature of FDA, FDA has advantages for class visualization from a theoretical perspective. Figures 5.19 and 5.20 show the scatter plot of PFP data in 2-D and 3-D Fisher spaces. It is obvious that both Figure 5.19 and Figure 5.20 better discriminate different classes than any plot based on PCA or PLS projections.

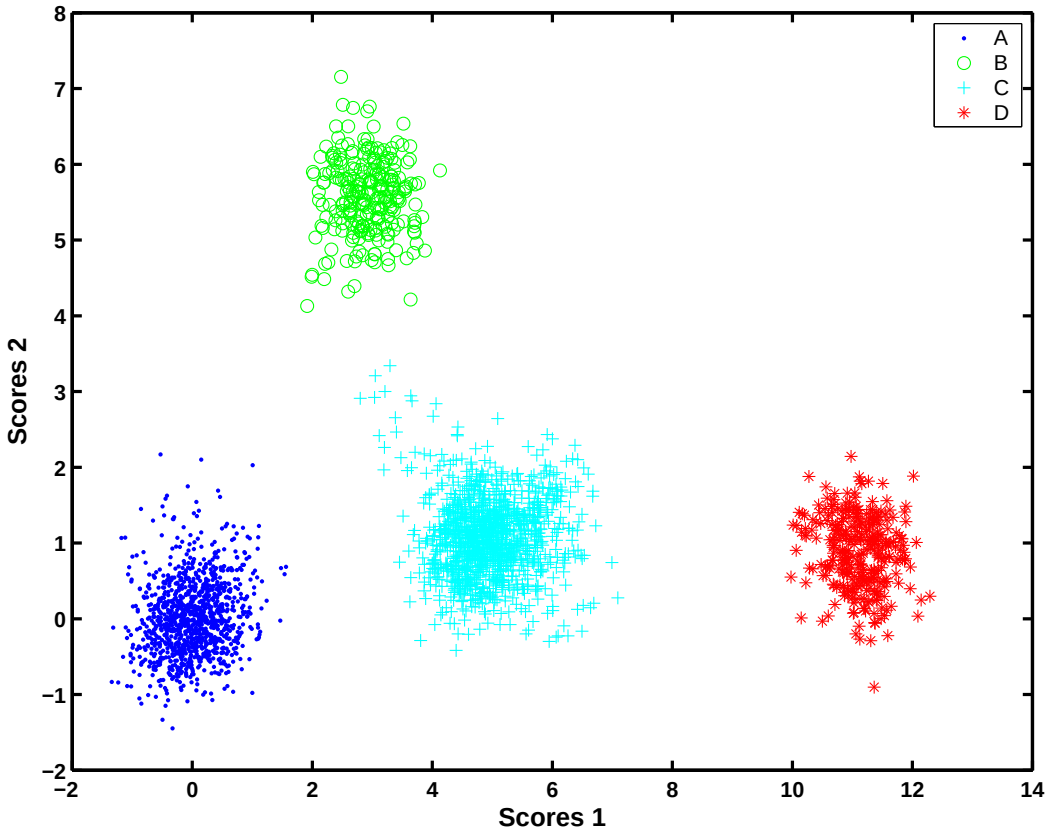


Figure 5.19: 2-D FDA score plot of the PFP data

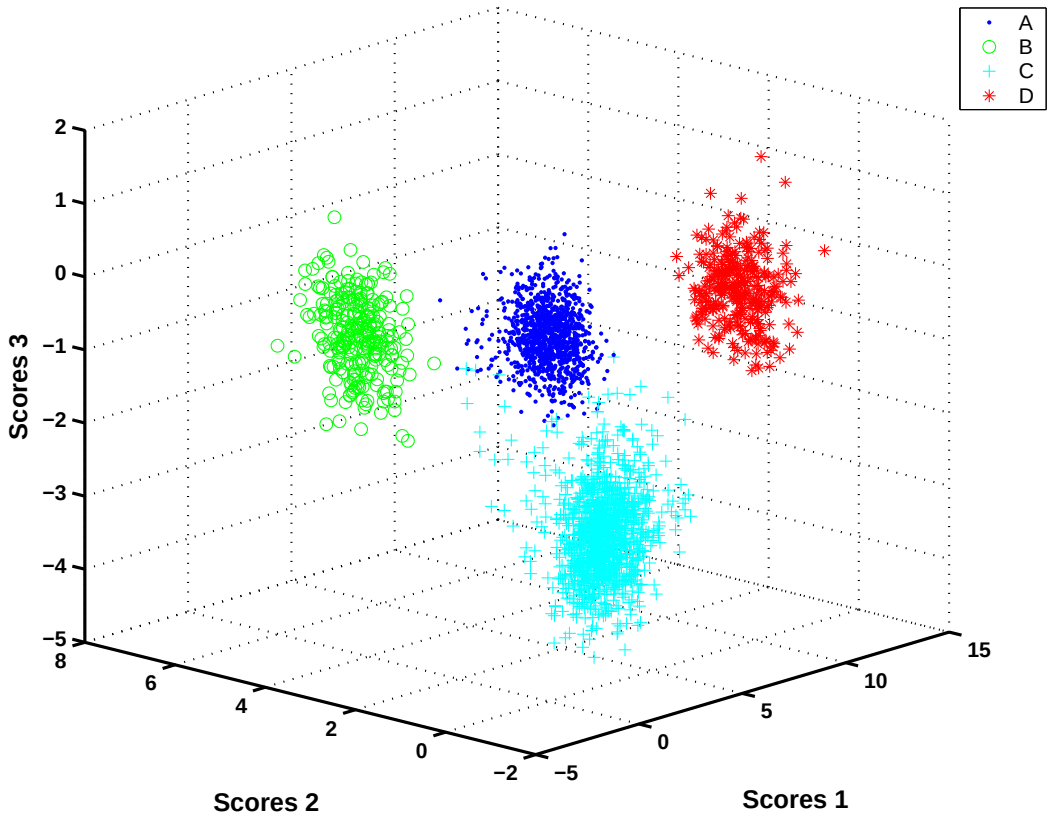


Figure 5.20: 3-D FDA score plot of the PFP data

5.4.4 Class-preserving projection (CPP)

CPP is closely related to FDA and its generalizations [21]. Instead of maximizing the ratio $\frac{\varphi^T S_b \varphi}{\varphi^T S_w \varphi}$ as FDA does, CPP maximizes $\varphi^T S_b \varphi$, i.e., CPP maximizes the distance between the projected means of different classes by considering the between-class scatter S_b only while ignoring the within-class scatter S_w. CPP has several advantages over FDA: First, FDA requires the solution of a generalized eigenvalue problem if S_w is singular, which can be computationally demanding for large data sets; CPP is dealing with an

eigenvalue problem. Second, FDA does not preserve the distance between class-means in the projection. CPP exactly preserves the distances between the class-means, that is, the distances between the projected means are exactly equal to the corresponding distances in the original high-dimensional space[1]. The score matrix T is obtained similarly as in FDA:

$$T = X\phi \tag{5.14}$$

Figures 5.21 and 5.22 show the 2-D and 3-D score plots based on CPP. Compared to FDA projection, although the distances between class means are "best" preserved, the compactness of each class and the clear separation among classes are sacrificed.

5.4.5 Support vector machines (SVM)

SVM is a binary classification method whose foundations have been developed by Vapnik [94]. SVM is developed to solve the classification problem, but they have been extended to the domain of regression problems [95]. In this work, we focus on its classification function where SVM takes labelled data from two classes as training data and generate a linear or nonlinear model for classifying new unlabelled data into one of those two classes. In the linearly separable case, which is the case studied in this work, SVM finds a separating hyperplane that maximize the margin of separation between the two

[1]Distances are exactly preserved only if there are three or less classes, otherwise, distances are preserved to the largest extent possible where the error due to projection is measured in the 2-norm or Frobenius norm, or any unitarily invariant norm [21].

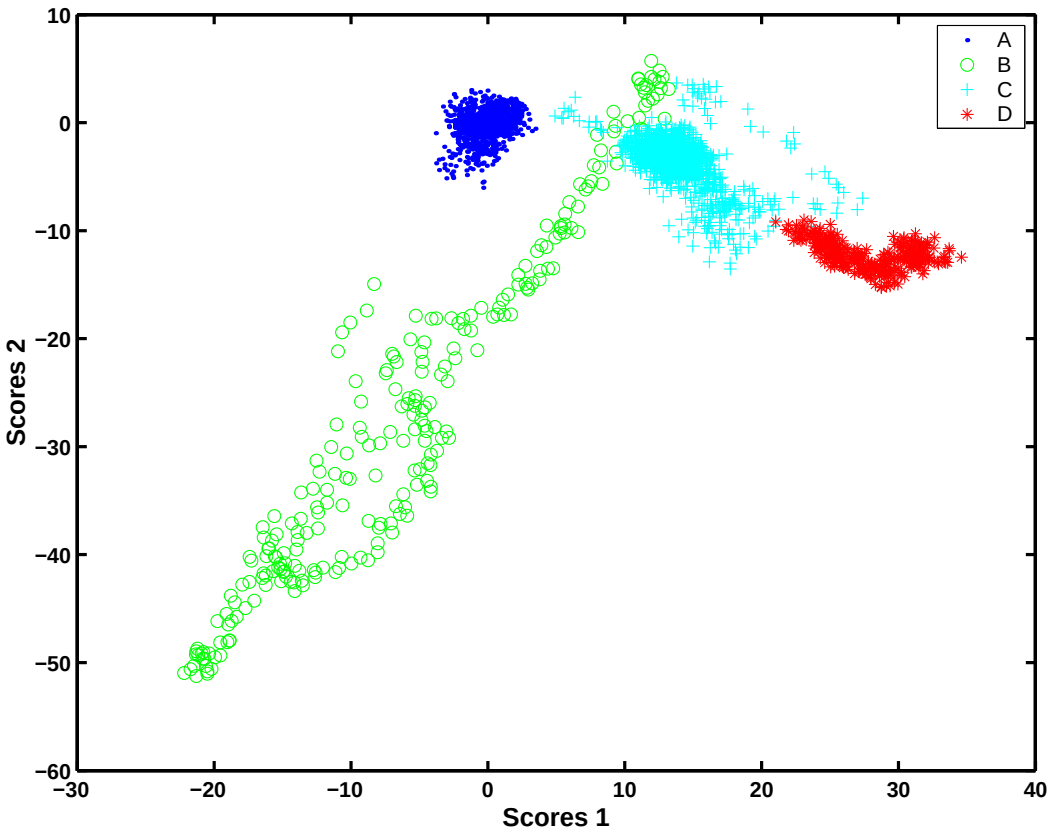


Figure 5.21: 2-D CPP score plot of the PFP data

classes. The algorithm finds a hyperplane which is the linear combination of the training data points, but most of the weights assigned to the data points are zeros. The points having nonzero weights are called support vectors S where $S = \{x_i | x_i \in X, \alpha_i \neq 0\}$. Consider the problem of separating the set of training vectors belonging to two separate classes,

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}, x \in \mathbf{R}^m, y \in \{-1, 1\} \tag{5.15}$$

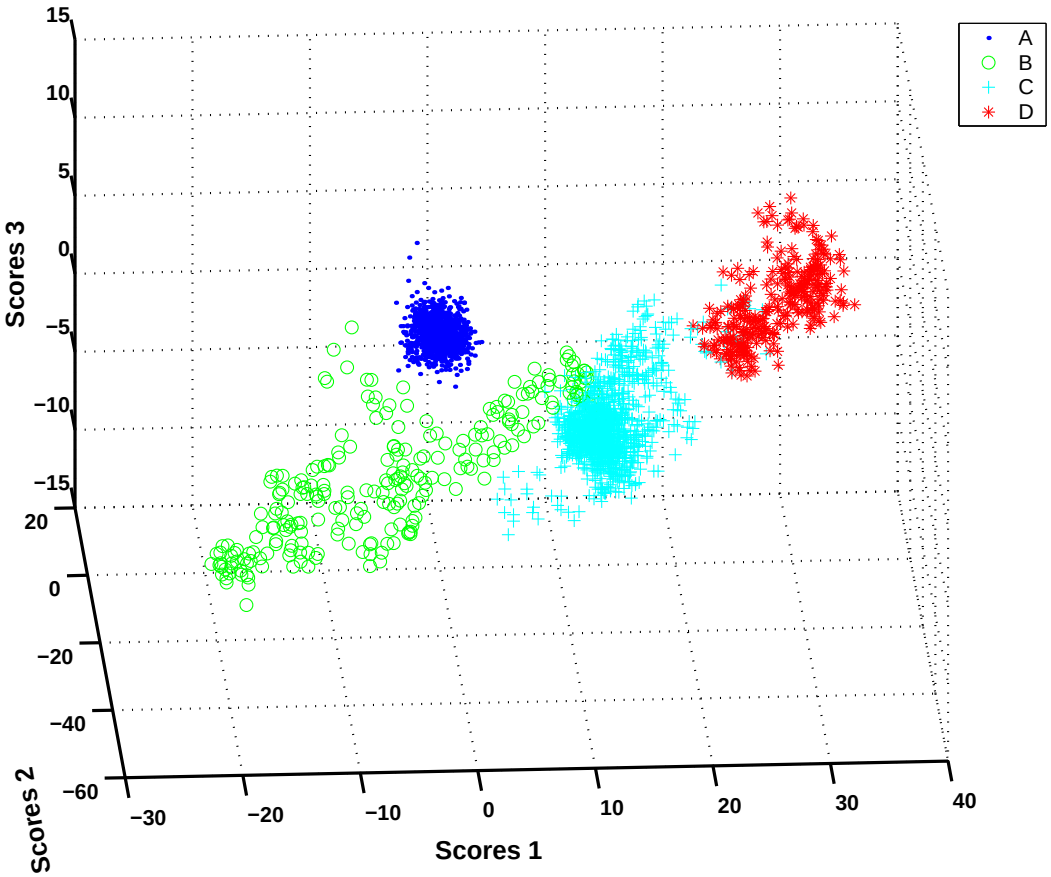


Figure 5.22: 3-D CPP score plot of the PFP data

where x_i is the m-dimensional data vector, y_i is the known class for data vector x_i. The separating hyperplane is defined as

$$\langle w, x \rangle + b = 0 \tag{5.16}$$

where

$$w = \sum_{i=1}^{s} (\alpha_i \cdot y_i) x_i, \tag{5.17}$$

where s is the number of support vectors, and α_i's are the support vector coefficients that maximize the margin of separation between the two classes. The classification for a new unlabelled point can be obtained from

$$f_{w,b}(x) = sign(w \cdot x + b) \tag{5.18}$$

For visualization purpose, we define the projection direction as the vector ω, which is normal to the separating hyperplane, and the transformed observations, i.e., the score vector:

$$t = X\omega \tag{5.19}$$

For 2-class case, in order to visualize the transformed observations in 2-D space, we obtain the first coordinates as:

$$t_1 = X\omega + b \tag{5.20}$$

The second coordinates are determined by applying SVM again to the observations other than support vectors S obtained previously for the first coordinates:

$$t_2 = X'\omega' + b' \tag{5.21}$$

where

$$X' = \{x_i | x_i \in X, x_i \notin S\} \tag{5.22}$$

Notice here t_1 and t_2 may not be orthogonal.

For multi-class case, since SVM is essentially a 2-class classifier, enhancements are required. A direct multi-class extension of SVM is considered in Vapnik, Crammer and Singer [18, 94], but usually the direct extension leads

to a very complex optimization problem and tedious computation. Therefore, multi-class problems are often solved by training several binary SVM classifiers and fusing the outputs of the classifiers to find the global classification decisions. Different coupling strategies are compared in Pöyhönen etc [78]. For K-class problem, $1/2K(K-1)$ 2-class classifiers are built. For visualization purpose, we propose a binary tree approach which does not require to build $1/2K(K-1)$ 2-class classifiers and coupling is not necessary. A schematic diagram is give in Figure 5.23 where observations consists of 4 classes (leaves) are projected onto 3 directions (t_1, t_2 and t_3) to optimally separate them. Alternatively, instead of separating individual class completely, i.e., each leaf consists of a single class, we propose another approach called cross-selection where each classifier separates all classes into two groups and different groups have part of their classes exchanged for each different classifier. Cross-selection approach does not create a tree structure, but rather creates a single level of clusters. In this way, the number of classifiers can be further reduced. Figure 5.24 shows a schematic diagram of the cross-selection approach where 4 classes are separated by 2 classifiers. Cross-selection approach may not be able to separate all classes completely even for the linear separable case. However, no matter how many classifiers we obtained, in order to visualize different classes in 2-D (3-D) space, only 2 (3) classifiers are actually used. Thus cross-selection approach with 2 or 3 classifiers has better chance to give good overall separation of classes than binary tree approach where more classifiers are needed to explicitly separate all classes but only 2 or 3 classifiers are utilized in the

visualization.

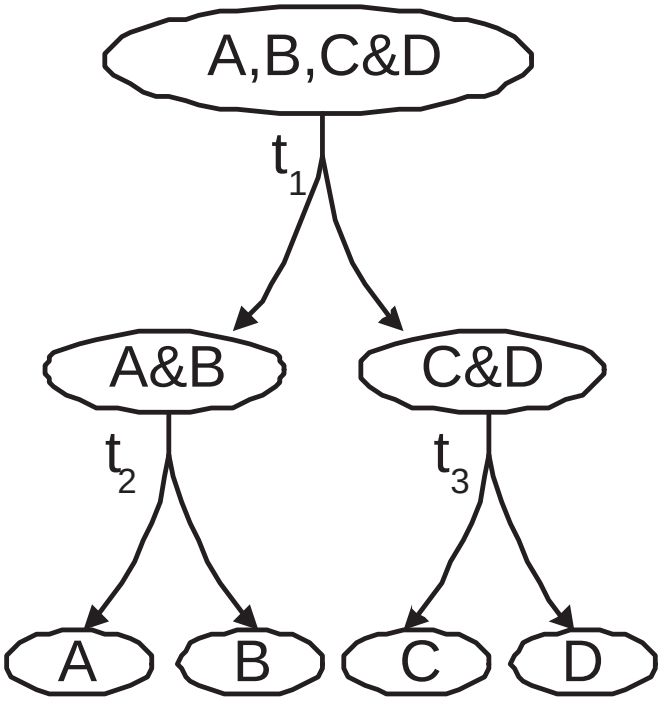


Figure 5.23: Schematic diagram of binary tree SVM approach

Unlike FDA based class visualization which consider both within-class and between-class scatter matrices, nor CPP which only consider within-class scatter matrix while try to preserve class distances in the original space, SVM classifier maximizes the distances between the boundaries of different classes. Figures 5.25 and 5.26 show the 2-D and 3-D visualization of the PFP data using the cross-selection approach and the binary tree approach. Both methods separate different classes very well while the cross-selection approach uses fewer classifiers (i.e., visualization dimensions) to achieve the same, if not better, separation.

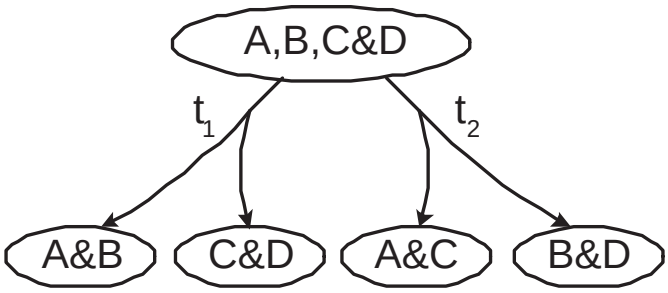


Figure 5.24: Schematic diagram of cross-selection SVM approach

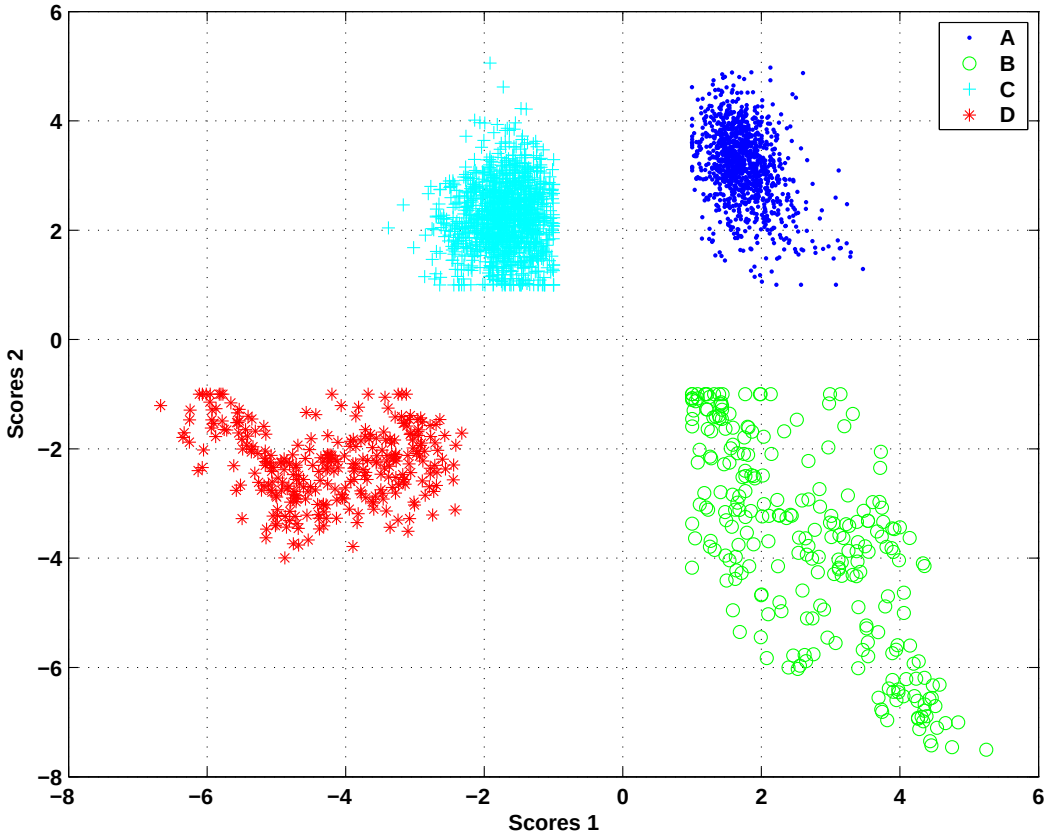


Figure 5.25: 2-D SVM score plot of the PFP data using the cross-selection approach

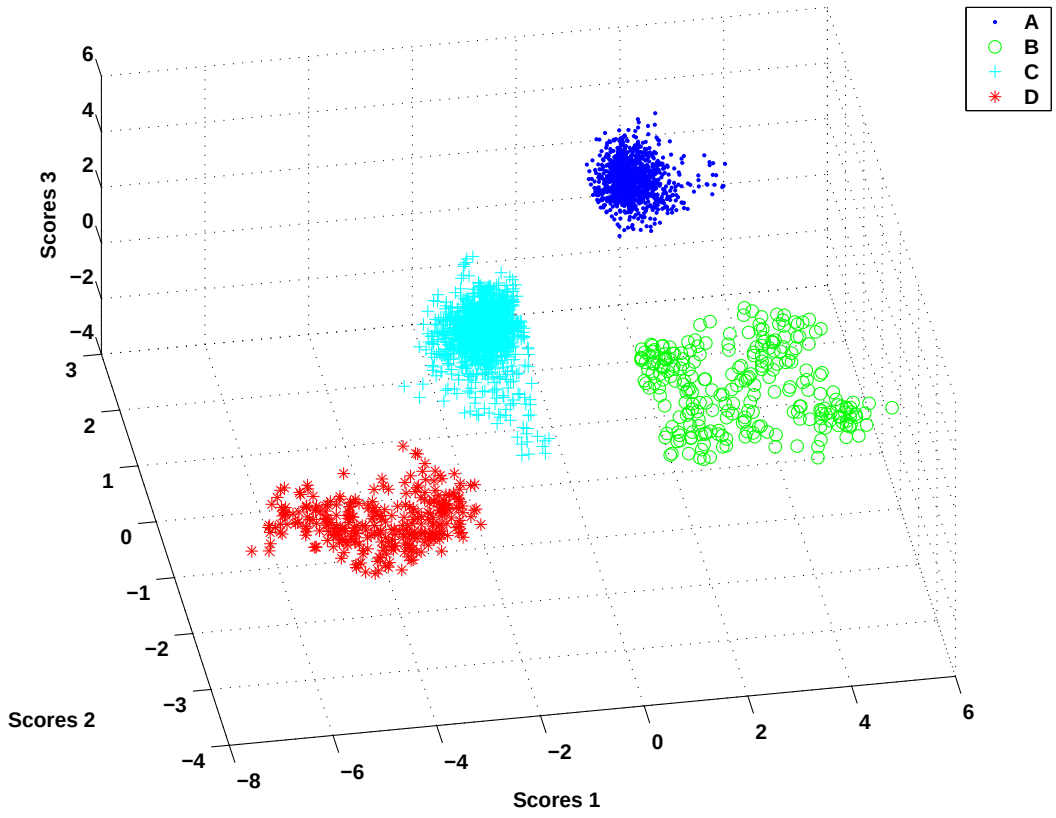


Figure 5.26: 3-D SVM score plot of the PFP data using the binary tree approach

5.5 Visualization of Process Dynamics in the Transformed Space

As we discussed in the previous section, process changes may occur not in any single dimension but in the combination of multiple dimensions, which makes the visualization of the process dynamics more difficult in the original space than in some transformed spaces. Besides, as large data sets become increasingly common nowadays, transformation of the high-dimensional data into lower-dimensional space would allow users to get clearer view and easier interpretation. In this section, we explore the visualization of process dynamics using PCA and FDA.

PCA based squared prediction error (SPE) charts and Hotelling's T^2 charts have been widely used in chemometrics to detect process dynamics changes, i.e., faults or operation mode changes. As an example, the PCA model is built for the TEP data based on the normal operation with 5 PC's. The time series plots of SPE and Hotelling's T^2 in Figure 5.27 successfully detected the process change occurred at 1500s. However, neither SPE nor T^2 tells us what the change is (process shift, drift, oscillation, or variance increase?). Since PCA transforms the original set of variables into a substantially smaller set of uncorrelated variables that represents most of the information in the original set of variables, we could think that the time series of the first several components would represent the process dynamics. If the PCA model is built based on the normal operation data, only the normal operation dynamics will be maximally represented by principal components. For example, the projec-

tions onto the first and second PC's (scores) shown in Figure 5.28 show the difference between the first half and the second half of the process. However, it does not represent the process dynamics very well. Alternatively, if the PCA model is built based on the data across the entire process, we would expect that the first several PC's would characterize the entire process. As we can see, Figure 5.29 captures the characteristics of the process: the process is at steady state during the first half of the process but starts to oscillate after that. This is consistent with Figures 5.5 and 5.9. As another example, Figure 5.30 shows the score plots of the PFP data corresponding to the projections onto the first three PC's where the PCA model is built based on the entire data set. If we compare this figure with Figure 5.31 where the variables with the ten largest variations are shown, it is easy to discover that scores 1 mimics the trends of most of those variables with mean shifts at different periods and different magnitudes, and scores 2 mimics the oscillations occurred in variables 25 and 28, and scores 3 is almost a copy of variable 74 with some effect from other variables, for example, variable 54. From these two examples, we see that the first few PC's usually capture the most significant dynamics of the process, which makes sense for plant engineers to monitor the first few PC's instead of all the original variables without missing the detection of process changes. For online monitoring, recursive PCA (RPCA) [70] can be implemented.

Notice that the class information is not required for above-mentioned PCA based process dynamics visualization method. If class information is given, FDA can also be used for process dynamics visualization. Figure 5.32

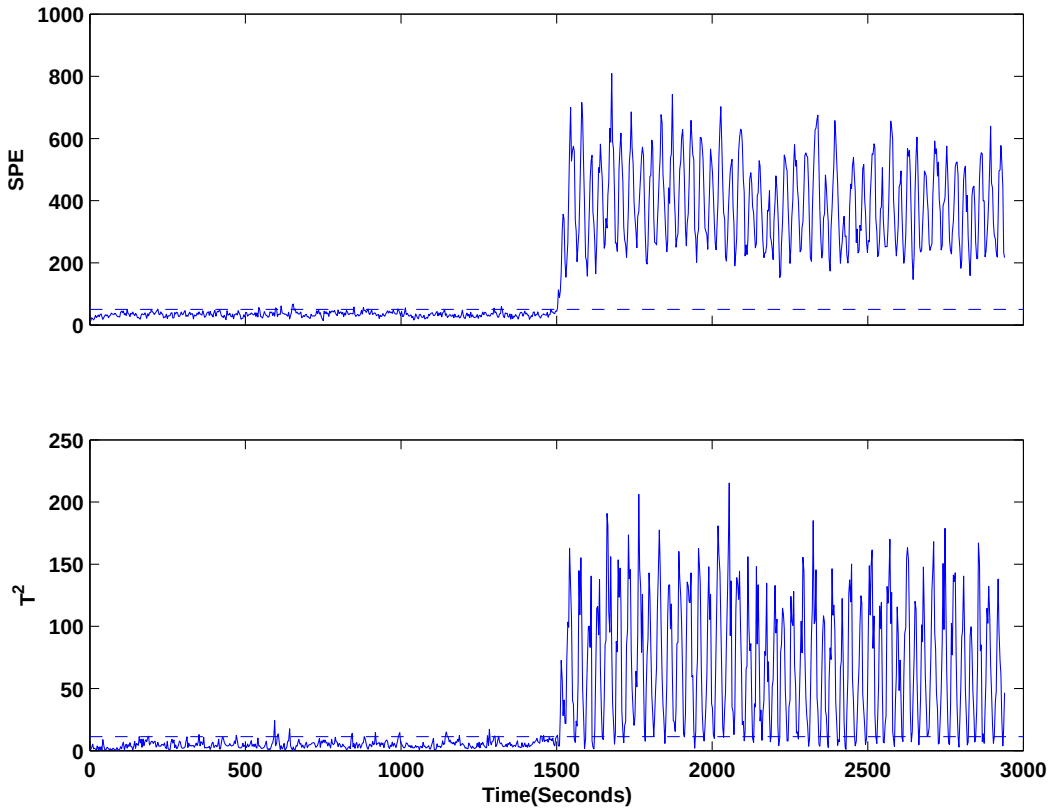


Figure 5.27: PCA SPE and T^2 plot of the TEP data (90% control limits are shown as dash lines)

shows FDA score plots of the PFP data. Compared to PCA score plots, however, it is a less representative picture of the process dynamics because FDA is developed for class discrimination by considering mean distances only but ignoring other characteristics such as variation changes, oscillations. For the TEP data, since there is no mean but variation change, FDA score plots as shown in Figure 5.33 fail to depict the process dynamics change.

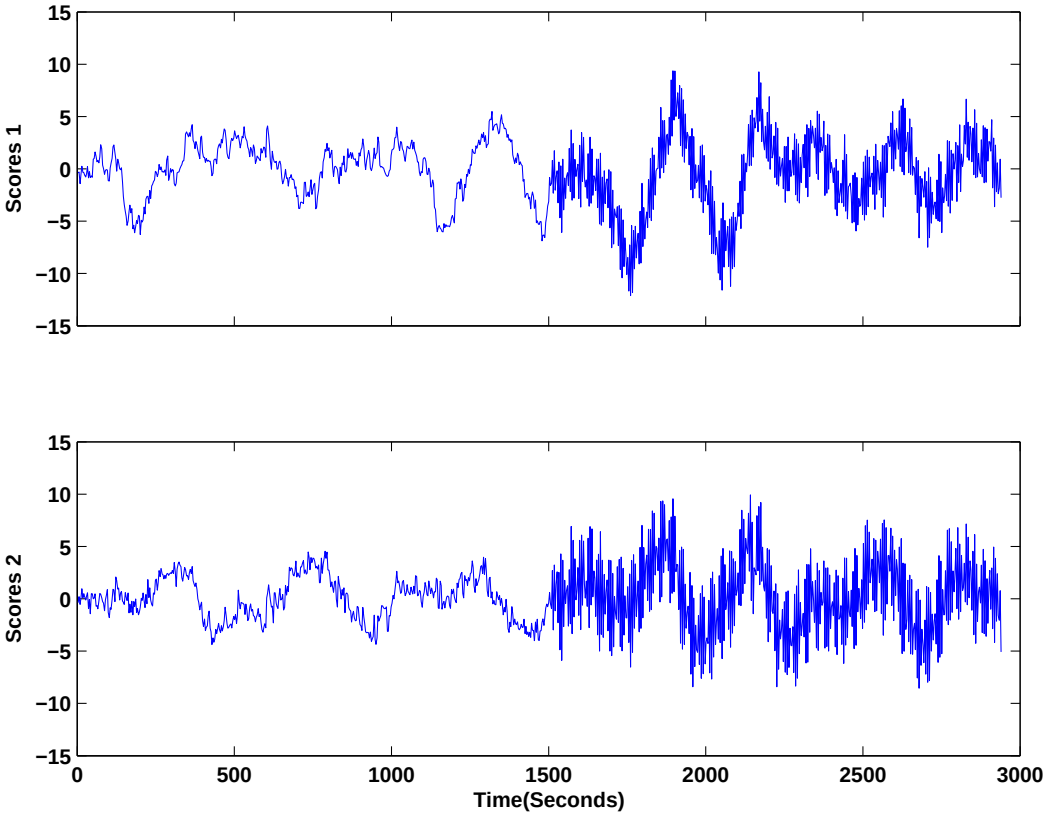


Figure 5.28: PCA score plots of the TEP data

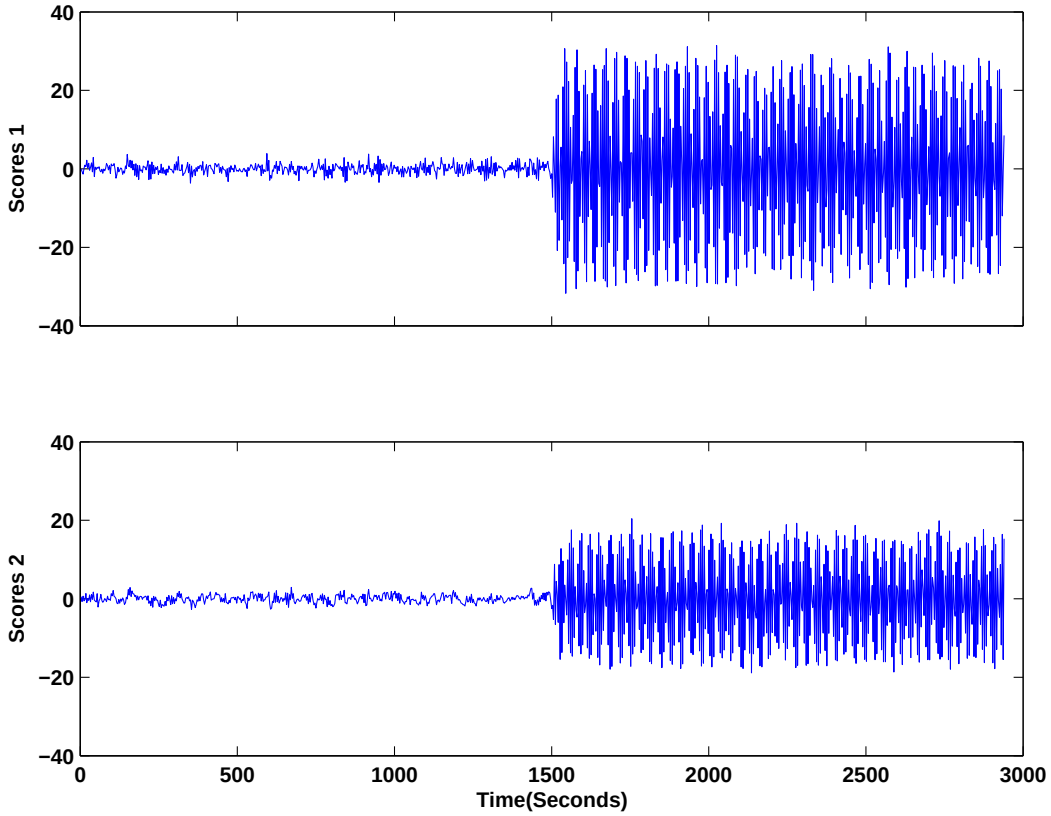


Figure 5.29: PCA score plots of the TEP data

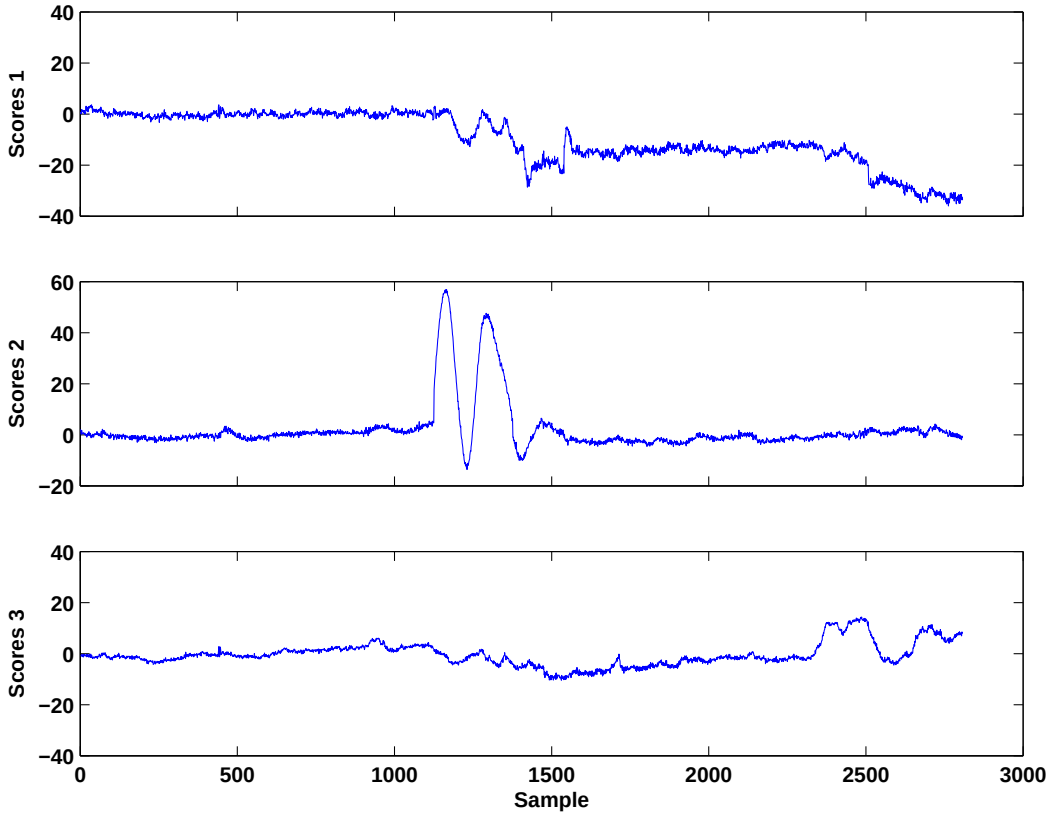


Figure 5.30: PCA scores plot of the PFP data

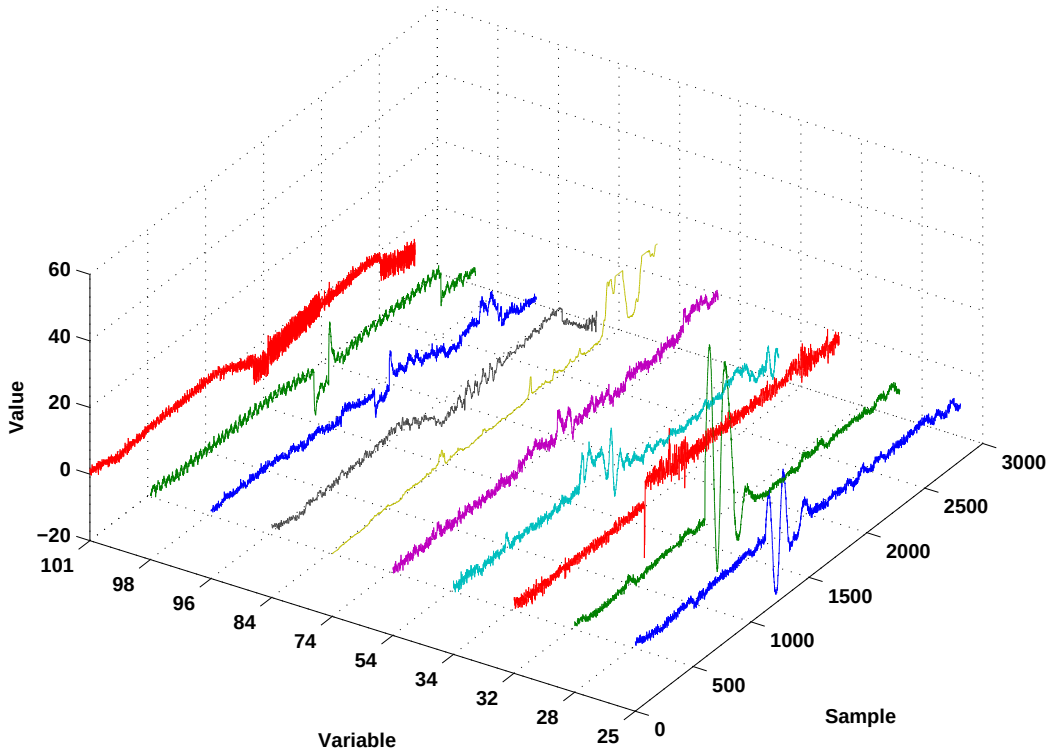


Figure 5.31: DPC plot of the PFP data with key variable identification

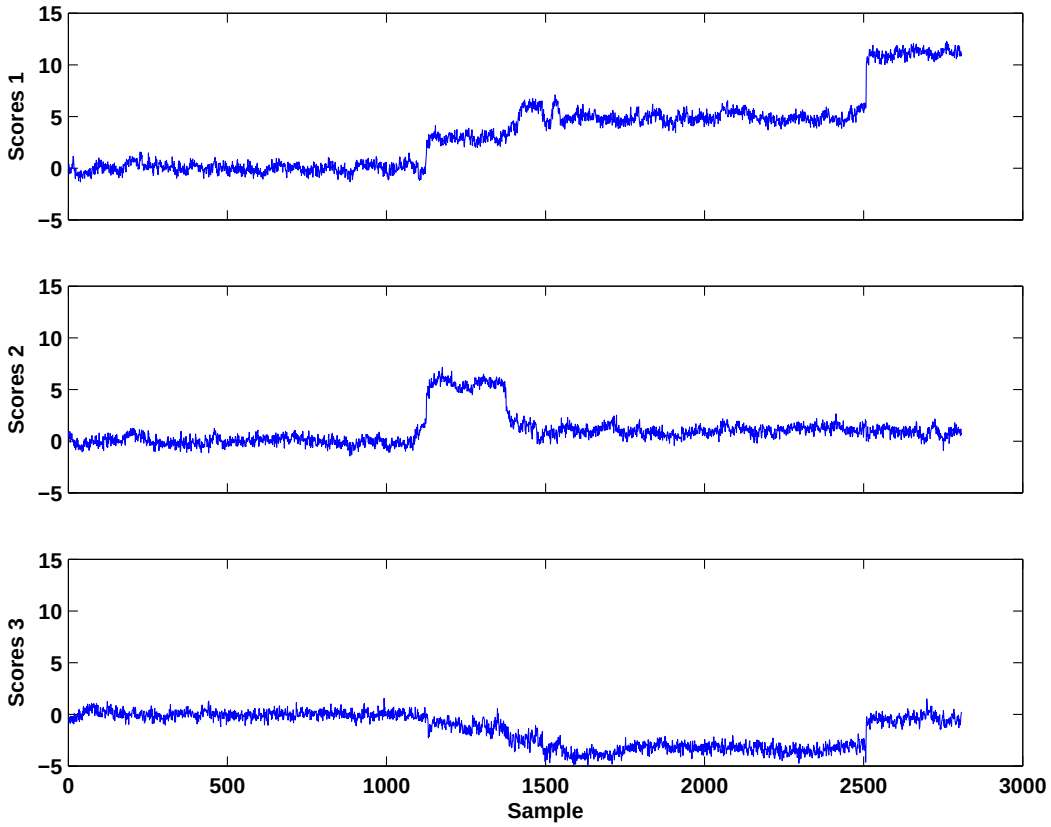


Figure 5.32: FDA scores plot of the PFP data

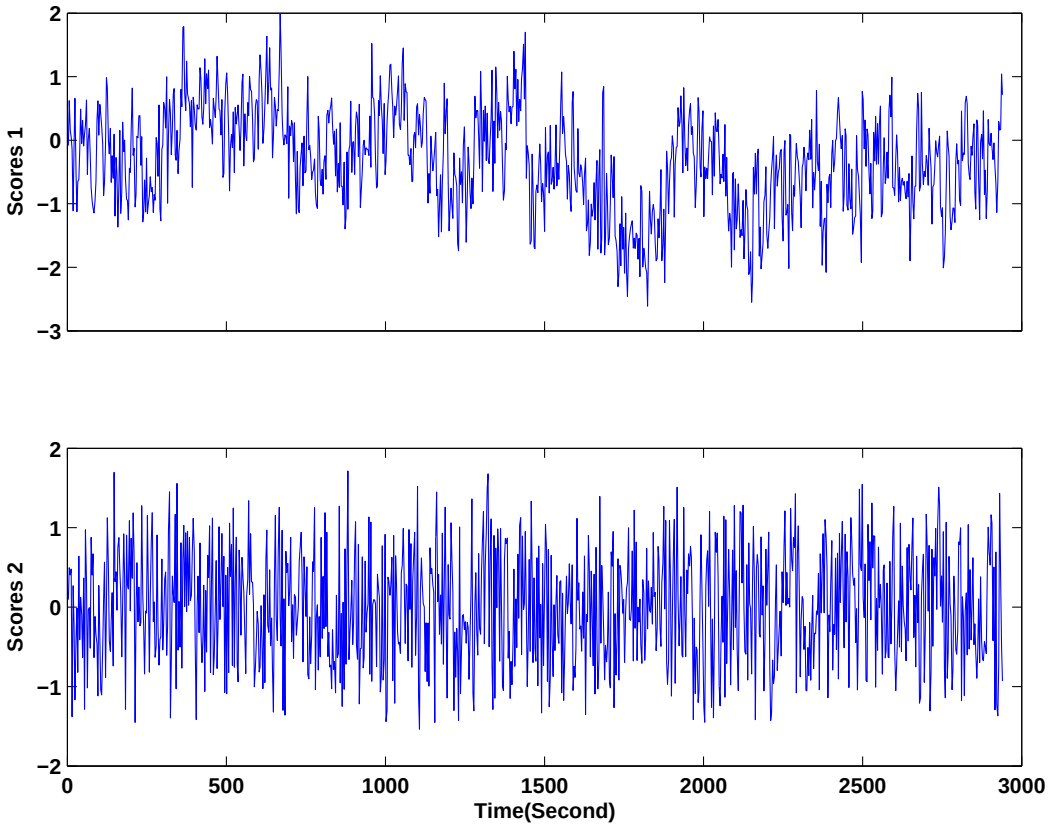


Figure 5.33: FDA scores plot of the TEP data

5.6 Conclusions

In this work, we have evaluated the most commonly used techniques and methods we proposed in large scale dynamic systems for the visualization of the static and dynamic properties in the original and transformed spaces. Scatter plots and parallel coordinates are most efficient in visualizing static properties for relatively small systems. Extruded parallel coordinates extends the capability of traditional parallel coordinates to visualization of process dynamics. Dynamic parallel coordinates is proposed to further improve the effectiveness in dynamic visualization. Contour plots are also explored and show some nice properties in the visualization of dynamic systems despite their drawbacks. For all visualization techniques applied in the original space, scaling, smoothing and key variable identification are important factors which affect visualization quality. We have shown that scaling based on the process specification or normal operation data provides much better results than auto-scale in terms of process dynamics visualization. We also propose key variable identification with threshold of variance or range as the criterion so that only variables with significant variance or range changes will be displayed. In addition to key variable identification, to tackle the clutter problem when visualizing large-scale data sets, we propose a general frame work of hierarchical visualization based on variable grouping. The effectiveness of this approach is illustrated in its application to a simulated chemical process. On class visualization of high dimensional data in transformed spaces, methods we proposed based on FDA and SVM show superior performance to PCA, PLS and CPP

in terms of compactness of each class and clear separation among different classes. We also have explored the visualization of process dynamics using PCA and FDA and have shown that the first few PC's usually capture the most significant dynamics of the process when PCA model is built based on the entire data set while FDA can only give a less representative picture of the process dynamics or even fails to depict the process dynamics if only variance changes are present in the process.

Chapter 6

Summary and Recommendations

6.1 Summary of Contributions

This dissertation presents several innovative techniques in industrial process modeling and monitoring. The major contributions of this dissertation are in the areas of industrial process modeling, fault diagnosis and industrial process visualization.

- In many complex processes, especially for semiconductor manufacturing processes, due to limited metrology information, feed back controls using simple linear empirical models are not sufficient to control the process tightly as complex interactions among manufacturing parameters are not represented accurately. To address this issue, some first principles models have been developed based on known physics and chemistry of the process to better capture the process characteristics. However, due to the complexity of the process, those first principles models usually contains tens or hundreds of differential equations and the required computation intensity makes the online application very difficult to meet the high-volume manufacturing needs. In this case some simplifications are necessary to make the model feasible for online control application,

while still preserves the dominant process characteristics. Therefore, the most important part of the modeling is to understand the physics and chemistry behind the process, and make reasonable simplifications and necessary transformations to build models accurate yet simple enough for online application. This motivated the LPCVD modeling in this dissertation and the major contributions are:

- A new thermal model is developed based on the energy balance analysis among furnace elements, which accurately predicts wafer temperatures using the furnace wall temperatures. The new model appears to be the first transformed linear model which captures the nonlinear relationship between the furnace wall temperature profile and the wafer temperature profile
- The model can be solved with a direct algorithm instead of iterative algorithms which are used in all existing thermal models. Since the direct algorithm is non-iterative, there is no convergence problem, nor local minima problem related to nonlinear optimization. In addition, the direct algorithm greatly reduces the computation effort.
- Configuration factors are calculated by a finite area to finite area method. This avoids numerical integration methods which are much more difficult to implement and require more computation time.
- Model sensitivity analysis are performed analytically to provide in-

sight into how cross-load wafer temperature profile can be affected by different heating elements and doors.

Even in the cases where first principles models are not feasible due to unknown or unmeasurable parameters, it is still vital to understand the underlying principles of the process. In this case, a hybrid model based on the mechanism of the process can be built to reproduce its behavior very well. This motivated the modeling of valve stiction and the major contributions are:

  - Two recently published valve stiction models are analyzed and their shortfalls are identified.
  - Based on the typical input-output behavior of a sticky valve, a new valve stiction model is developed which has much simpler structure and more straightforward logic compared to other existing models. The new model can be easily implemented to simulate a sticky valve as part of a control loop to help understand the valve stiction phenomenon and its impact on the performance of the control loop.

- Another trend in process industry is that massive amount of data become more and more common, and the question is how to extract useful information from the data and make use of it. One important area is fault detection and diagnosis and many statistical analysis methods have been developed in this area. In order to make better use of these methods, differences between different methods should be understood, and

new methods can be developed by taking advantage of these differences. This motivated the proposal of fault diagnosis based on fault directions defined by FDA. Because FDA makes use of available information from fault data, it has the advantage in fault diagnosis compared to PCA, where only the information from normal data is used for model building. Major contribution in this area are summarized below:

- A new fault diagnosis method using fault directions in Fisher discriminant analysis is developed. The developed method shows superior capability for fault diagnosis to the contribution plots method based on PCA.
- The fault direction is defined for the first time as the Fisher direction which optimally discriminates fault data from normal data. This direction best characterizes the effect of the fault relative to the normal data and the weights in the fault direction are used to generate the contribution plot for fault diagnosis.
- A new process monitoring method is proposed which consists of data pre-analysis, fault visualization and fault diagnosis and the method is illustrated using an industrial film process.

This dissertation also makes contributions to fault diagnosis of oscillating control loops, where a curve fitting method is proposed to detect valve stiction based on the study of the characteristics of a sticky valve:

- The inconsistency of Horch's first method is theoretically analyzed for the first time and illustrated by a simulated example. A curve fitting method for detecting valve stiction is proposed which is applicable to both self-regulating and integrating processes and its theoretical analysis is presented. The proposed method shows superior performance to other existing methods in both simulated and industrial examples. The proposed method is industrial-oriented and it works fully automatically without user interactions.

- In the area of process monitoring, statistical process monitoring (SPM) has become one of the most active research areas in the last decade. However, the visualization of the static and dynamic behavior of large complex processes has been difficult and largely unsolved issue and not much effort has appeared in this area. In this dissertation, an extensive study on visualization techniques is conducted and major contributions are summarized below:

  - Several commonly used visualization techniques, usually applied to relatively small static systems, are evaluated in the context of large dynamic systems. Their advantages and disadvantages are discussed.

  - Dynamic parallel coordinates (DPC) is proposed to visualize large data sets and capture their dynamic characteristics.

- Several factors affecting the quality of visualization are discussed and variable grouping is introduced to reduce clutter in handling large data sets.
- Hierarchical visualization scheme is proposed to provide a general framework for visualization and exploration of large multivariate data sets.
- Two approaches based on SVM, i.e., the binary-tree approach and the cross-selection approach, are proposed for high dimension class visualization.

6.2 Suggestions for Future Work

Future research directions which deserve further investigation are summarized in this section.

- The optimization of the cross-load wafer temperature profile based on the proposed model can be further investigated.
- The application of the proposed thermal model to control should be investigated further.
- Kinetic modeling can be investigated and incorporated with the proposed thermal model to predict the film thickness across the wafer load.
- The proposed thermal model is an one-dimensional model. Extension to a 2-D model while still keeping its simple structure and light computation

load could be a challenging and rewarding problem.

- The fault diagnosis method using pair-wise FDA proposed in this work only points out which variables contribute to the fault. We do not know what happened to that variable and there could be different faults to the same variable. If the historical data for the known faults are available, it is desirable to classify the new fault based on the known fault types.

- Robust fault diagnosis which is independent of the magnitude, time duration and direction of the fault and the plant operating point deserves further investigation. The robust fault diagnosis should be able to handle a large number of noisy variables as well.

- It is of interest to develop new methods based on pattern classification techniques for fault detection, fault identification and diagnosis, fault estimation and fault reconstruction.

- The operation region classification could be another future direction.

Bibliography

[1] H. Albazzaz, X. Z. Wang, and F. Marhoon. Multidimensional visualization for process historical data analysis: a comparative study with multivariate statistical process control. *J. Proc. Cont.*, 15:285–294, 2005.

[2] K.J. Astrom. Assessment of achieviable performance of simple feedback loops. *Int. J. Adaptive Control and Signal Processing*, 5(1):3–19, 1991.

[3] C. Azzaro and J. P. Couderc. Thermal modeling of tubular horizontal hot-wall low pressure chemical vapor deposition reactors. *The Chemical Engineering Journal*, 57:39–52, 1995.

[4] T. A. Badgwell, T. F. Edgar, I. Trachtenberg, G. Yetter, J. K. Elliott, and R. L. Anderson. In situ measurement of wafer temperatures in a low pressure chemical vapor deposition furnace. *IEEE Transactions on Semiconductor Manufacturing*, 6(1):65–71, 1993.

[5] T. A. Badgwell, I. Trachtenberg, and T. F. Edgar. Modeling the wafer temperature profile in a multiwafer LPCVD furnace. *J. Electrochem. Soc.*, 141(1):161–172, 1994.

[6] Thomas A. Badgwell. *Modeling and Optimization of Multiwafer Low*

Pressure Chemical Vapor Deposition Reactors. PhD thesis, The University of Texas at Austin, 1992.

[7] A. Benveniste, M. Basseville, and G. Moustakides. The asymptotic local approach to change detection and model validation. *IEEE Trans. Auto. Cont.*, 32(7):583–592, July 1987.

[8] W. L. Bialkowski. Dreams vs. reality: A view from both sides of the gap. *Pulp and Paper Canada*, 94(11):19–27, 1993.

[9] Edited by James Moyne, Enrique del Castillo, and Arnon Max Hurwitz. *Run-to-Run Control in Semiconductor Manufacturing*. CRC Press LLC, Boca Raton, FL, 2001.

[10] Stephen A. Campbell. *The Science and Engineering of Microelectronic Fabrication*. Oxford University Press, Inc., second edition, 2001.

[11] H. Chernoff. Using faces to represent points in k-dimensional space. *Journal of the American Statistical Association*, 68:361–368, 1973.

[12] L. H. Chiang, E. L. Russell, and R. D. Braatz. Fault diagnosis and fisher discriminant analysis, discriminant partial least squares, and principal component analysis. *Chemometrics Intell. Lab. Syst.*, 50:243–252, 2000.

[13] L. H. Chiang, E. L. Russell, and R. D. Braatz. *Fault Detection and Diagnosis in Industrial Systems*. Springer, London, 2001.

[14] M. A. A. S. Choudhury, N. F. Thornhill, and S. L. Shah. A data-driven model for valve stiction. In *IFAC Symposium on Advanced Control of Chemical Processes (ADCHEM)*, Hong Kong, January 11-14 2004.

[15] M. A. A. S. Choudhury, N. F. Thornhill, and S. L. Shah. Modelling valve stiction. *Control Engineering Practice*, 13:641–658, 2005.

[16] A. K. Conlin, E. B. Martin, and A. J. Morris. Confidence limits for contribution plots. *J. Chemomotrics*, 14:725–736, 2000.

[17] D. G. Coronell and K. F. Jensen. A monte carlo simulation study of radiation heat transfer in the multiwafer LPCVD reactor. *J. Electrochem. Soc.*, 141(2):496–501, 1994.

[18] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

[19] R. Deibert. Model based fault detection of valves in flow control loops. In *IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes – SAFEPROCESS '94*, pages 445–450, Espoo, Finland, 1994.

[20] L. Desborough, P. Nordh, and R. Miller. Control system reliability: process out of control. *Industrial Computing*, 2(8):52–55, 2001.

[21] I. S. Dhillon, D. S. Modha, and W. S. Spangler. Class visualization of high-dimensional data with applications. *Computational Statistics and Data Analysis*, 28(1):59–90, 2002.

[22] R. Dunia and S. J. Qin. Subspace approach to multidimensional fault identification and reconstruction. *AIChE J.*, 44:1813–1831, 1998.

[23] G. H. Dunteman. *Principal components analysis*. Sage Publications, Inc., Newbury Park, CA, 1989.

[24] D. B. Ender. Process control performance: Not as good as you think. *Control Engineering*, Sept.:180–190, 1993.

[25] S. Fienberg. Graphical methods in statistics. *The American Statistician*, 33:165–178, 1979.

[26] K. Forsman and A. Stattin. A new criterion for detecting oscillations in control loops. In *European Control Conference*, Karlsruhe, Germany, 1999.

[27] P.M. Frank. Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy — a survey and some new results. *Automatica*, 26:459–474, 1990.

[28] P. Geladi and B. R. Kowalski. Partial least-squares regresion: A tutorial. *Analytica Chimica Acta*, 185:1–17, 1986.

[29] P. Geladi, M. Manley, and T. Lestander. Scatter plotting in multivariate data analysis. *J. Chemomotrics*, 17:503–511, 2003.

[30] J. Gerry and M. Ruel. How to measure and combat valve stiction online. In *ISA International Fall Conference*, Houston, TX, Sept. 10–13 2001.

[31] J. Gertler. Survey of model-based failure detection and isolation in complex plants. *IEEE Cont. Sys. Mag.*, 12:3–11, 1988.

[32] M. Gevers. A personal view on the development of system identification. In *Proceedings of 13th IFAC symposium on System Identification*, pages 773–784, Rotterdam, Netherlands, 2003.

[33] G. C. Goodwin, S. F. Graebe, and M. E. Salgado. *Control System Design*. Prentice Hall, 2001.

[34] E. Groller, H. Loffelmann, and R. Wegenkittl. Visualization of dynamical systems. *Future Generation Computer Systems*, 15(1):75–86, 1999.

[35] T. Hagglund. A control-loop performance monitor. *Control Eng. Practice*, 3(11):1543–1551, 1995.

[36] D.M. Hawkins. The detection of errors in multivariate data using principal components. *J. Amer. Stat. Asso.*, 69:340–344, 1974.

[37] Q. He and M. Pottmann. Detection of valve stiction using curve fitting. Internal Report, Process Dynamics and Control, DuPont Engineering, August, 2003.

[38] Q. He and S. J. Qin. Multivariate visualization in statistical process monitoring. to be submitted to Journal of Chemometrics.

[39] Q. He and S. J. Qin. Multivariate visualization in data analysis for process operations. In *AIChE Annual Conference*, Austin, TX, November 2004.

[40] Q. He, S. J. Qin, and A. J. Toprac. A new thermal model for the hot-wall low pressure chemical vapor deposition. In *AIChE Annual Conference*, Indianapolis, IN, November 2002.

[41] Q. He, S. J. Qin, and A. J. Toprac. Computationally efficient modeling of wafer temperatures in a low pressure chemical vapor deposition furnace. *IEEE Transactions on Semiconductor Manufacturing*, 16(2):342–350, 2003.

[42] Q. He, S. J. Qin, and A. J. Toprac. Computationally efficient modeling of wafer temperatures in a lpcvd furnace. In *SPIE Advanced Process Control and Automation*, Santa Clara, CA, February 2003.

[43] Q. He, S. J. Qin, and A. J. Toprac. CVD modeling. In *Texas-Wisconsin Modeling and Control Consortium Spring Meeting*, Austin, TX, February 2003.

[44] Q. He, J. Wang, M. Pottmann, and S. J. Qin. A curve fitting method for detecting valve stiction in oscillating control loops. to be submitted to Journal of Process Control.

[45] Q. He, J. Wang, M. Pottmann, and S. J. Qin. A curve fitting method for detecting valve stiction in oscillatory control loops. In *Texas-Wisconsin Modeling and Control Consortium Spring Meeting*, Austin, TX, February 2005.

[46] Q. He, J. Wang, and S. J. Qin. Fault diagnosis using fault directions in fisher discriminant analysis. In *AIChE Annual Conference*, San Francisco, November 2003.

[47] Q. He, J. Wang, and S. J. Qin. Fault diagnosis and visualization using fisher discriminant analysis. In *Texas-Wisconsin Modeling and Control Consortium Spring Meeting*, Austin, TX, February 2004.

[48] Q. He, J. Wang, and S. J. Qin. A new fault diagnosis method using fault directions in fisher discriminant analysis. *AIChE J.*, 51(2):555–571, 2005.

[49] D. W. Hess, K. F. Jensen, and T. J. Anderson. Chemical vapor deposition: a chemical engineering perspective. *Reviews in Chem. Eng.*, 3:97–186, 1985.

[50] S. Hirasawa, S. Kieda, T. Watanabe, T. Torii, T. Takagaki, and T. Uchino. Temperature distribution in semiconductor wafers heated in a vertical diffusion furnace. *IEEE Transactions on Semiconductor Manufacturing*, 6(3):226–232, 1993.

[51] S. Hirasawa and T. Takagaki. In *Proceedings of the 1989 National Heat Transfer Conference*, 1989.

[52] A. Horch. A simple method for detection of stiction in control valves. *Control Engineering Practice*, 7:1221–1231, 1999.

[53] A. Horch. *Condition monitoring of control loops.* PhD thesis, Royal Institute of Technology, Stockholm, Sweden, December 2000.

[54] A. Horch and A. J. Isaksson. A method for detection of stiction in control valves. In *IFAC Workshop on On-Line Fault Detection and Supervision in The Chemical Process Industry*, Lyon, France, 1998.

[55] W. G. Houf, J. F. Grcar, and W. G. Breiland. A model for low pressure chemical vapor deposition in a hot-wall tubular reactor. *Materials Science and Engineering*, B17:163–171, 1993.

[56] John R. Howell. *A catalog of radiation configuration factors.* McGraw-Hill, Inc., 1982.

[57] S. M. Hu. Temperature distribution and stresses in circular wafers in a row during radiative cooling. *J. Appl. Phys.*, 40(11):4413–4423, 1969.

[58] A. Inselberg and B. Dimsdale. Parallel coordinates: A tool for visualizing multidimensional geometry. In *Proceedings of the 1st IEEE Conference on Visualization (Vis '90)*, pages 361–378, 1990.

[59] R. Isermann. Process fault detection based on modeling and estimation methods - a survey. *Automatica*, 20:387–404, 1984.

[60] J. E. Jackson and G. Mudholkar. Control procedures for residuals associated with principal component analysis. *Technometrics*, 21:341–349, 1979.

[61] K. H. Johansson. The quadruple-tank process: a multivariable laboratory process with an adjustable zero. *IEEE Trans. Cont. Sys. Tech.*, 8(3):456–465, 2000.

[62] P. A. Jokinen. Visualization of multivariate processes using principal component analysis and nonlinear inverse modelling. *Decision Support Systems*, 11:53–65, 1994.

[63] I.T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, London, 1986.

[64] M. Kano, H. Maruta, H. Kugemoto, and K. Shimizu. Practical model and detection algorithm for valve stiction. In *Proceedings of the Seventh IFAC-DYCOPS Symposium*, Boston, USA, July 2004.

[65] I. K. Kim and W. S. Kim. Theoretical analysis of wafer temperature dynamics in a low pressure chemical vapor deposition reactor. *International Journal of Heat and Mass Transfer*, 42:4131–4142, 1999.

[66] T. Kourti and J. F. MacGregor. Multivariate SPC methods for monitoring and diagnosing of process performance. In *Proceedings of PSE*, pages 739–746, 1994.

[67] T. Kourti and J. F. MacGregor. Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometrics Intell. Lab. Syst.*, 28:3–21, 1995.

[68] T. Kourti and J. F. MacGregor. Multivariate SPC methods for process and product monitoring. *J. Qual. Tech.*, 28:409–428, 1996.

[69] J.V. Kresta, J. F. MacGregor, and T. E. Marlin. Multivariate statistical monitoring of processes. *Can. J. Chem. Eng.*, 69(1):35–47, 1991.

[70] W. Li, H. Yue, S. Valle-Cervantes, and S.J. Qin. Recursive PCA for adaptive process monitoring. *J. Proc. Cont.*, 10:471 –486, 2000.

[71] J. F. MacGregor. Statistical process control of multivariate processes. In *Preprints IFAC ADCHEM*, 1994.

[72] J. F. MacGregor, C. Jaeckle, C. Kiparissides, and M. Koutoudi. Process monitoring and diagnosis by multiblock PLS methods. *AIChE J.*, 40:826–828, 1994.

[73] I. Matsuba, K. Mokuya, K. Matsumoto, and A. Yoshinaka. Mathematical model of temperature distribution in wafers in a furnace for

semiconductor fabrication processes. In *Proceedings of the 4th International Conference on the Numerical Analysis of Semiconductor Devices and Integrated Circuits (NASECODE IV)*, pages 405–410, 1985.

[74] T. Miao and D. E. Seborg. Automatic detection of excessively oscillatory feedback control loops. In *Proc. IEEE Intl. Conf. on Control Applications*, pages 359–364, Kohala Coast - Island of Hawaii, August 1999.

[75] P. Miller, R.E. Swanson, and C.E. Heckler. Contribution plots: A missing link in multivariable quality control. *Appl. Math. and Comp. Sci.*, 8(4):775–792, 1998.

[76] P. Nomikos. Statistical monitoring of batch processes. In *Preprints of Joint Statistical Meeting*, Anaheim, CA, August 1997.

[77] K. S. Park, M. Choi, H. J. Cho, and J. D. Chung. Analysis of radiative heat transfer and mass transfer during multiwafer low-pressure chemical vapor deposition. *J. Electrochem. Soc.*, 147(12):4554–4561, 2000.

[78] S. Poyhonen:arkkio:jover:hyotyniemi. Coupling pairwise support vector machines for fault classification. *Control Engineering Practice*, in press, 2004.

[79] S. J. Qin, S. Valle-Cervantes, and M. Piovoso. On unifying multiblock analysis with applications to decentralized process monitoring. *J. Chemomotrics*, 15:715–742, 2001.

[80] S. Joe Qin. Statistical process monitoring: Basics and beyond. *J. Chemomotrics*, 17:480–502, 2003.

[81] S.J. Qin. Control performance monitoring – a review and assessment. *Computers and Chemical Engineering*, 23:178–186, 1998.

[82] A. Raich and A. Cinar. Statistical process monitoring and disturbance diagnosis in multivariate continuous processes. *AIChE J.*, 42:995–1009, 1996.

[83] A. M. Rinaldi, S. Carra, M. Rampoldi, M. C. Martignoni, and M. Masi. Lpcvd vertical furnace optimization for undoped polysilicon film deposition. *J. de Physique IV*, 9:189–196, 1999.

[84] P.E. Hart R.O. Duda and D. G. Stork. *Pattern Classification.* John Wiley & Sons, New York, 2nd edition, 2001.

[85] M. Rossi and C. Scali. Automatic detection of stiction in actuators: A technique to reduce the number of uncertain cases. In *IFAC-DYCOPS - 7th International Conference*, Cambridge, MA, USA, 2004.

[86] T. Sato. Spectral emissivity of silicon. *Japanese Journal of Applied Physics*, 6(3):339–347, 1967.

[87] B. J. Van Schravendijk, W. L. De Koning, and W. C. Nuijen. Modeling and control of the wafer temperatures in a diffusion furnace. *J. Appl. Phys.*, 61(4):1620–1627, 1987.

[88] Robert Siegel and John R. Howell. *Thermal Radiation Heat Transfer*. Taylor & Francis, third edition, 1992.

[89] A. Singhal and T. I. Salsbury. A simple method for detecting valve stiction in oscillating control loops. *J. Proc. Cont.*, 15:371–382, 2005.

[90] O. Taha, G. A. Dumont, and M. S. Davies. Detection and diagnosis of oscillations in control loops. In *35th IEEE Conference on Decision and Control*, pages 2432–2437, Kobe, Japan, 1996.

[91] M. A. Tavel and E. W. Hearn. An interactive computer simulation of heating and cooling a row of silicon wafers. *J. Electrochem. Soc.*, 135(5):1266–1271, 1988.

[92] N. F. Thornhill and T. Hagglund. Detection and dagnosis of oscillation in control loops. *Control Eng. Practice*, 5(10):1343–1354, 1997.

[93] H. Tong and C. M. Crowe. Detection of gross errors in data reconciliation by principal component analysis. *AIChE J.*, 41:1712–1722, 1995.

[94] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 2nd edition, 2000.

[95] V. N. Vapnik, S. E. Golowich, and A. Smola. Support vector method for function approximation, regression estimation and signal processing. In *Neural Information Processing Systems*, pages 281–287, Cambridge, MA, 1997.

[96] H. De Waard and W. L. De Koning. Optimal control of the wafer temperatures in diffusion/LPCVD reactor. *Automatica*, 28(2):243–253, 1992.

[97] A. Wallen. Valve diagnostics and automatic tuning. In *Proceedings of the American Control Conference*, pages 2930–2934, Albuquerque, New Mexico, 1997.

[98] E. J. Wegman. Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, 85(411):664–675, 1990.

[99] J. A. Westerhuis, S.P. Gurden, and A.K. Smilde. Generalized contribution plots in multivariate statistical process monitoring. *Chemometrics Intell. Lab. Syst.*, 51:95–114, 2000.

[100] A. E. Widmer and W. Rehwald. Thermoplastic deformation of silicon wafers. *J. Electrochem. Soc.*, 133:2403–2409, Nov 1986.

[101] B.M. Wise and N.B. Gallagher. The process chemometrics approach to process monitoring and fault detection. *J. Proc. Cont.*, 6:329–348, 1996.

[102] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics Intell. Lab. Syst.*, **2**:37–52, 1987.

[103] J. Yang, M. O. Ward, E. A. Rundensteiner, and S. Huang. Visual hierarchical dimension reduction for exploration of high dimensional datasets.

In *Eurographics/IEEE TCVG Symposium on Visualization*, pages 19–28, 2003.

[104] S. Yoon and J.F. MacGregor. Fault diagnosis with multivariate statistical models, part I: using steady state fault signatures. *J. Proc. Cont.*, 11:387–400, 2001.

[105] H. Yue and S. Joe Qin. Reconstruction based fault identification using a combined index. *Ind. Eng. Chem. Res.*, 40:4403–4414, 2001.

[106] H. Yue and S.J. Qin. Fault reconstruction and identification for industrial processes. In *AIChE Annual Meeting*, Miami, FL, Nov. 1998.

[107] Q. J. Nottingham D. F. Cook C. W. Zobel. Visualization of multivariate data with radial plots using SAS. *Computers and Industrial Engineering*, 41:17–35, 2001.

Vita

Qinghua He, the second son of Benyu He and Dehui Jiang, was born in Guanghan, Sichuan, P.R. China on July 13, 1974. He received his high school diploma in July 1991 from Guanhan high school in Guanghan, Sichuan. Mr. He entered Tsinghua University, Beijing in the fall of 1991, and graduated in July 1996, with the Bachelor of Engineering in Chemical Engineering. After working as a research scientist at Tsinghua University for three years, Mr. He entered graduate school at the Pennsylvania State University at University Park in August 1999, and transferred to the University of Texas at Austin in May 2000, where he earned the Master of Science in Chemical Engineering in December 2002. He was admitted to Ph.D. candidacy at the University of Texas at Austin in September 2003. Mr. He joined Advanced Micro Devices as a process development engineer in May 2004.

Permanent address: 10901 Beachmont Lane
Austin, Texas 78739

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.